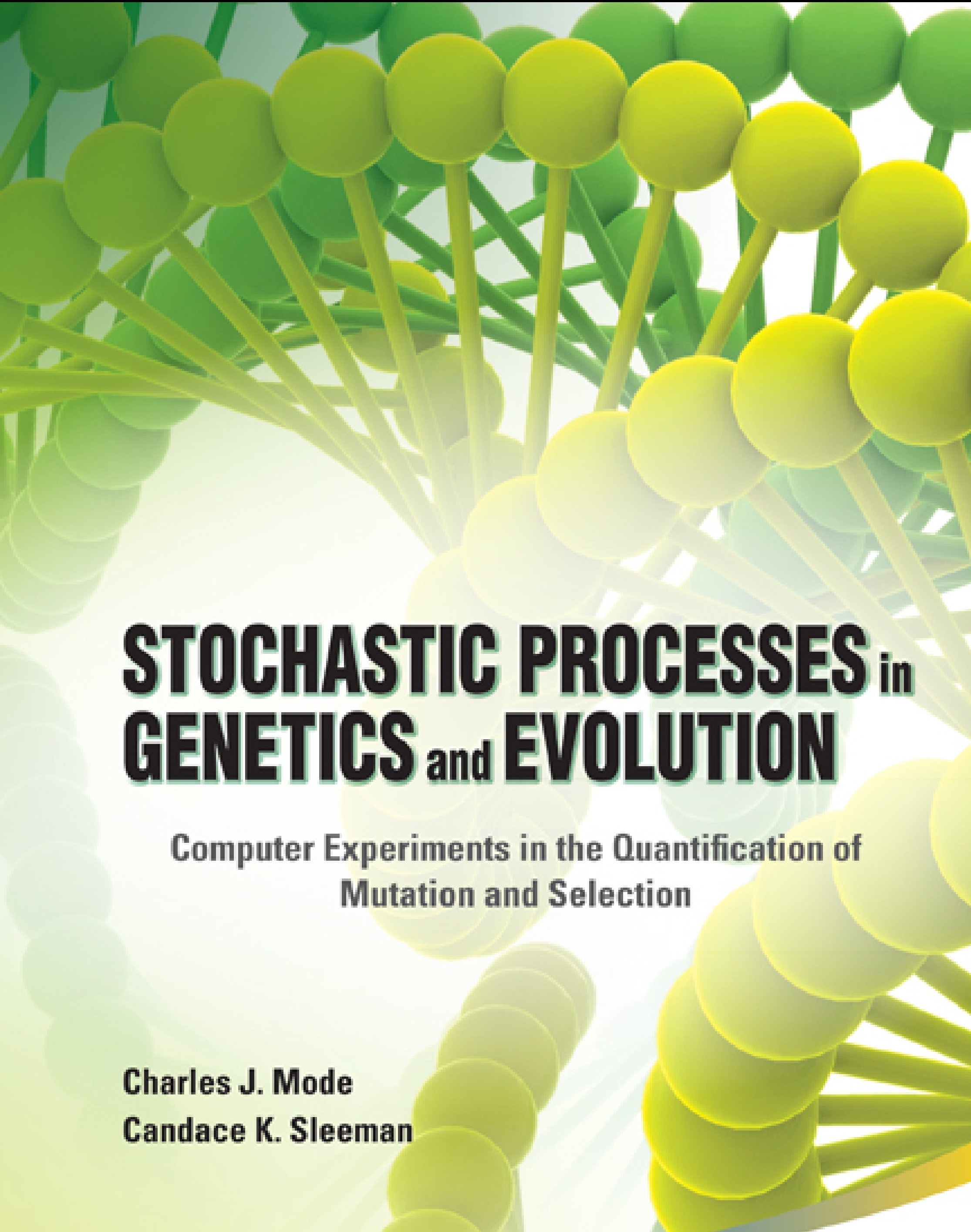

STOCHASTIC PROCESSES in GENETICS and EVOLUTION

Computer Experiments in the Quantification of Mutation and Selection

Charles J. Mode
Candace K. Sleeman

World Scientific

STOCHASTIC PROCESSES in GENETICS and EVOLUTION

Computer Experiments in the Quantification of Mutation and Selection

This page intentionally left blank

STOCHASTIC PROCESSES in GENETICS and EVOLUTION

Computer Experiments in the Quantification of Mutation and Selection

Charles J. Mode
Drexel University, USA

Candace K. Sleeman
NAVTEQ Corporation, USA

World Scientific

NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI

Published by

World Scientific Publishing Co. Pte. Ltd.
5 Toh Tuck Link, Singapore 596224
USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601
UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

STOCHASTIC PROCESSES IN GENETICS AND EVOLUTION
Computer Experiments in the Quantification of Mutation and Selection

ISBN-13 978-981-4350-67-9
ISBN-10 981-4350-67-2

Printed in Singapore.

Dedication

To the memory of my wife, Eleanore L. Perdelwitz Mode and my parents, Karl Charles and Fanny E. Hansen Mode

To the memory of my father, Dr. Richard A. Sleeman

Prologue

At the outset it should be stated that this is not a book on phylogenetics in which millions of years of evolution and relationships among existing species are often under consideration. However, in chapters 6,7 and 8 stochastic models of nucleotide substitutions, which may be applied in research on phylogenetics, are reviewed and some useful extensions are suggested that accommodate nucleotide substitutions at a large number of sites of a DNA molecule rather than a single site or codons with three sites that are characteristic of most of the models introduced to the literature 3 to 4 decades ago. In contrast to research in phylogenetics, the main thrust of this book is to provide methods for simulating the stochastic evolution of single species during short periods of evolutionary time consisting of 10,000 to 200,000 years, but in some models time is expressed on a scale of generations.

A common theme of the computer simulation experiments reported in this book is the evolution of a population stemming from a small founder population. Of particular interest in these simulation experiments is the informative statistical summarization of a sample of Monte Carlo waiting times until a new beneficial mutation arises and become predominant in a population. Another distinguishing feature of this book is that all Monte Carlo simulation models are rooted in stochastic processes, and, in each case, an attempt has been made to present the mathematics in sufficient detail so that if an investigator were interested, he would, in principle, be possible to write software in a programming language of his choosing and duplicate any computer experiment reported in this book. The programming language used throughout this book was APL 2000, which is an international language that is popular among a minority of people who like the succinctness with which complex code may be written. Unfortunately, this programming language is not as popular as C^{++} and other languages. As mathematics is a universal international language however, it is hoped that by inspecting the mathematics underlying a model, an investigator will be able to write software to implement any model discussed in a programming language of his choosing.

The following paragraphs of this prologue are devoted to suggestions that will be helpful to readers who wish to read this book thoroughly or merely skim through it or even skip some chapters to obtain an overall impression of the contents of the book. It is hoped that the modular nature of this book by topics will expedite this exploratory process.

Chapter 1 is devoted to an axiomatic treatment of probability, which will be useful in setting the stage for the chapters that follow. A central theme of this chapter is the concept of a finite probability space, which encompasses a sample space of outcomes of a conceptual experiment, a collection of events or subsets of the sample space and the definition a probability function defined on the class of events with certain properties. Random variables are then defined within the context of a probability space and the binomial, multinomial and Poisson distributions are derived and are applied extensively throughout the book. For those readers who are not comfortable with the axiomatic approach to probability, it will be sufficient to grasp the ideas underlying the binomial, multinomial and Poisson distribution. In this connection, a study of the many examples from Mendelian genetics involving applications of these distributions will be very helpful.

Chapter 2 is devoted the parameterization of the gametic distribution with respect to a large number of linked Mendelian loci or markers such as single nucleotide polymorphisms, $SNPs$, on molecules of DNA. After some suggestions for assembling data bases to study genetic recombination at the molecular level, a method is developed for parameterizing the gametic distribution in terms of recombination probabilities for some arbitrary number $N \geq 2$ of linked loci. In the closing section of this chapter, suggestions are made as to how the ideas developed in the foregoing sections can be applied to pedigrees in which linked markers are under consideration.

Chapter 3 is devoted large random mating diploid populations with no mutation or selection and the principal objective of the chapter is to develop a mathematical structure that can accommodate a large number of linked loci with a finite but arbitrary number of alleles at each locus so that convergence to a linkage equilibrium in such a population may be studied. This chapter begins with a classical account of convergence to linkage equilibrium for the case of two loci with two alleles at each locus, which may be found in many text books of population genetics. This result is then extended to the case of a finite but arbitrary number of alleles at each locus and then finally to the general case of multiple loci mentioned above. Much of the content of chapter 3 is based on results by H. Geirenger, which were published in 1944. Among other things, this theory is based on a elegant application of set theory, and if this is discomforting to a reader, he can rest with the knowledge that this theory will not be used in subsequent chapters of the book. However, when genetic recombination is again encountered in chapter 14, the case of two linked loci discussed in chapters 2 and 3 will be applied to the case of two linked markers at the molecular level.

Chapter 4 is devoted to a presentation the Wright-Fisher process within the context of finite absorbing Markov chains in which applications of matrix theory are useful in reducing the structure of this process to simple terms that are familiar to anyone with a working knowledge of the theory of finite matrices. In particular, formulas of a set of conditional absorption probabilities are derived such that if the process starts in given transient state, the conditional probability that it is absorbed or the process in terminated in some particular absorbing state are expressed as elements of matrices. Furthermore, giving that the process terminates in some absorbing state, formulas for the conditional expectations and variances of the waiting times to absorption are derived. A general formula for the quasi-stationary distribution of a finite absorbing Markov chain are also derived, which will be useful in connections with branching processes introduced in subsequent chapters. Any mention of diffusion approximations to Wright-Fisher process have deliberately been avoided, because, for the most part, this book is devoted to computer intensive methods. In this chapter, Wright-Fisher processes with respect to a single autosomal locus with two alleles are the principal foci of attention and both the neutral case and the cases of mutation and selection as characterized within the Wright-Fisher paradigm in terms of probabilities. A class of Wright-Fisher processes with a state space such that all states communicate with each other was also included in this chapter.

Chapter 5 is devoted for the most part to Wright-Fisher process with multiple alleles at a single autosomal locus. As through trial and error it was found that matrix formulas derived in chapter 4 tend to become numerically unstable when the size of a Markov transition matrix exceeds about 1000 × 1000, it became necessary to use Monte Carlo simulation methods for dealing with process based on multiple alleles which usually entails the use of very large transition matrices. Fortunately, by using Monte Carlo simulation methods, problems with numerical instabilities can be avoided and the evolution of populations consisting one million or more individuals may be studied for thousands of generations on desk top computers, where the execution times for each experiment with 100 or more replications may be accomplished within ten to twenty minutes. The application of Monte Carlo simulation methods in this chapter also necessitated a treatment of some theories underlying the computer generation of random numbers as well as the description of statistical methods for the informative summarization of Monte Carlo simulation data, which were subsequently used in some of following chapters of this book.

Chapters 6, 7 and 8 are devoted to the mutational process of nucleotide substitutions. In chapter 6, an overview of the fundamentals underlying Markov jump processes in continuous time with finite state spaces is given, and there is also a brief discussion of a probability space on which this class of stochastic processes lives. It is also shown that the exponential matrix function, provides a solution to the Kolmogorov differential equations for the case of finite state spaces in a general case. Following this overview, specific examples of nucleotide substitution models based on this class of processes are reviewed. For the more simple models, a symbolic computation engine was used to derive explicit formulas for the exponential matrix, which provides explicit functions for the matrix of current state probabilities of the process. For more complex examples, numerical forms of the exponential matrix were computed, given assumed numerical values of the parameters. In this chapter there are also discussions that illuminate the process of nucleotide substitutions if it does indeed follow the laws of evolution implicit in the theory of Markov jump processes in continuous time. For example, given values of the rate parameters of the process, one could estimate the expectation of the time a particular nucleotide spends at one site of a DNA molecule until a transition to another nucleotide at this site occurs.

When one considers the problem of extending a nucleotide substitution model from one site of a DNA molecule to many sites, a simple and straight forward approach to the problem would be that of assuming nucleotide transitions among sites occur independently with the same rate matrix for each site. But, it has been proposed in the literature, that substitutions may occur at different rate among the sites of a molecule and that sites may not evolve independently in a probabilistic sense. When there are different rate parameters for each site, the problem of dealing with many parameters also arises. Consequently, it becomes necessary to devise a structure that circumvents problems of dealing with many parameters and at the same time to formulate a model such that evolution of substitutions among sites are dependent in some sense. Such a formulation was considered in chapter 7.

Briefly, in this approach it was assumed that the rate matrices at many sites were realizations of a stationary stochastic process, depending on only a few parameters and constructed on the basis of a consistent family of finite dimensional distribution functions, where consistency is defined in papers and books dealing with the foundations of probability. Given a realization of this process, it was assumed that substitutions at different sites were governed by conditionally independent Markov jump processes with four states. When averaged over realizations of the rate process to obtain unconditional distributions of the processes of among the sites, however, independence among the sites no longer holds. It also suggested in this chapter that the rate process could be constructed form simple Gaussian processes based on first or second order autoregressive processes. The range of the sample functions of these processes is the real line $(-\infty, \infty)$, but this set can always be mapped into the set of positive real numbers in $(0, \infty)$, which is the rate space for Markov jump processes in continuous time.

Chapter 8 is devoted to a software implementation of the stochastic structure developed in chapter 7 along with computer simulation experiments on nucleotide substitutions in the D loop of the human mitochondrial genome, which consists of 1,120 base sites. The classification of Haplo groups in existing human population of the world is based on about 22 $SNPs$ that occur mostly in the D loop. Given an initial DNA sequence consisting of 1,120 bases, a computer simulation experiment was run until 22 mutation were accumulated and each of these experiments were replicated 50 times. The evolutionary time taken to complete each replication varied, because the rates governing the Markov jump processes with random. When classifying Haplo groups, two complicating issues are often mentioned. One is the problem of back mutation and the other is the problem of parallel mutations. By focusing attention within each replication, it was possible to write software to estimate the frequency of back mutation, in which the initial nucleotide at a given site mutates to some other base and then back mutates to the initial base at that site. It was also possible to estimate the frequency of parallel mutations among the 50 replication that could be interpreted as separate evolving human populations. A mutation is said to be parallel if same mutation at a particular site occurs in more than one replication. Back mutations complicate the classification of Haplo groups, because individuals who ought to belong to a group because of shared ancestry are excluded due to the absence of a mutation. Parallel mutations complicate the problem of identifying a descendant of a founder of population based on whether he or she has a particular mutation, which could have been present in the actual founder of the group or is, on the other hand, the descendant of an individual who migrated into the group with a parallel mutation. Although attention was focused only on the D loop of the human mitochondrial genome, it is now known that the structure implemented in chapter could easily be extended to the entire mitochondrial genome that consists of about 16,000 bases or even large regions of DNA with larger number of bases.

Up until chapter 9, no stochastic structure had been entertained within which a range of demographic factors could be taken into account in simulating the evolution of human and other populations. As a first step towards correcting this omission, an outline of the one-type Galton-Watson branching process was given in chapter 9, which evolves on a time scale of discrete time generations. One of the serious limitations of this process, as well a the class of branching processes in general, which has been recognized for a long time, is that there are only two possibilities; namely, either the population becomes extinct or it increases without bound. To correct this limitation, the idea of a self regulating branching process was introduced. In this formulation, it is supposed that the probability that an individual in one generation survives to produce offspring in the next generation is a function of total population size with a parameter that indicates the population size that must be reached before extensive mortality occurs. It should be stated that this formulation is not simply a rework of the famous deterministic logistic model of population growth that has attracted much attention, because for some points in its parameter space, iterates of its defining equation become chaotic.

One of the innovational aspects of this chapter is that by using a one-type self regulating branching processes to simulate genealogies, it is possible to estimate the distribution of the number of generations back in time to the most recent common ancestor of any two randomly selected individuals in some current generation. This approach is an alternative way of looking at the problem of coalescence, which is mentioned in this chapter. A second innovation of self regulating branching processes was that it was possible to embed a deteministic model in a stochastic process, and it was shown by examples that, at some points in its parameter space, iterates of its defining equation become chaotic. Moreover, in computer simulation experiments, it was possible to compare the performance of statistically summarized sample of Monte Carlo realizations of the stochastic process with the chaotic embedded deterministic model, see chapter 9 for details.

In order for Darwinian selection to occur, a population must contain two or more types. Consequently, in chapter 10 is devoted experiments in the quantification of mutation and selection within a framework of self regulating multitype branching processes, which evolve on a discrete time scale expressed in terms of generations. In these experiments, two components of natural selection were taken into account in the formulation. One was a measure of reproductive success as expressed in terms of the expected number of total offspring contributed by each individual of a given type to the next generation of an evolving population. The other component was the ability of individuals of each type to survive to produce offspring of the next generation. This ability was characterized in terms of parameters, depending on type, which provided a threshold value such that when total population size exceeded this threshold, the probability that an individual survived to reproduce was reduced. Only three types or genotypes were considered in the experiments reported in this chapter, and, mutations among the types were described in terms probabilities of one type mutating to another type per generation. In these experiments, selection was quantified by assigning numerical values to expectations of total number of offspring contributed the next generation to member of each type as well as the threshold parameters in the survival function. Mutations were quantified by assigning numerical values to the probabilities governing mutations among the types in each generation. It was also possible for this class of branching processes, to embed deterministic vector-valued non-linear difference equations in the process, so that trajectories of the evolution of the population could be compared with trajectories based on statistically summarized Monte Carlo simulation data.

One of the most innovative experiments reported in this chapter was that if one type had a threshold parameter that was greater than that of other types and the measures of reproductive success of the types were equal, then it was shown that the type with the greatest threshold parameter eventually rose to predominance in the population within a few thousand generations even though this type had arisen by mutation during the evolution of a small founder population in which the beneficial mutation was not present. Interestingly, in this experiment it was possible to study the rise of this mutation in a stochastically evolving population as well as a population evolving according to the embedded deteministic model. As illustrated in the Monte Carlo simulation data, there were high levels of stochasticity in the beginning generations during the emergence of the mutation that was not, as expected, present in the trajectories computed using the embedded deterministic model. If an investigator confines his attention to the deterministic model in the studying the evolution of a multitype population, the presence of high level of stochasticity that exists during the emergence of a mutation would be missed.

Diploid populations of humans as well as those of other species are comprised of two sexes, females and males, who form sexual partnerships which usually contribute offspring to the next generation. When formulating stochastic models describing the evolution of such populations, it is necessary to include a module that can accommodate mating systems and the possibility that sexual selection may occur during mating process. Chapter 11 is devoted to the formulation and the computer implementation of stochastic models accommodating two sexes in populations that evolve in discrete time generations. In this class of models, the components of natural selection include the type of mating system, random or non-random, preferences of both females and males with respect to the phenotype in selecting their prospective sexual partners, measures of reproductive success by couple types as classified by genotype or phenotype and a probability that each individual female or male survives to contribute offspring to the next generation. Like all other stochastic population processes considered in this book, the stochastic process formulated in this chapter is also self-regulating. For the most part, the underlying genetics used in this chapter was confined to one autosomal locus with two alleles at each locus so that for both sexes only three genotypes were under consideration. Probabilities that each allele may mutate to the other per generation were also included in this formulation.

Perhaps one of most interesting computer experiments reported in this chapter was a case in which the only component of natural selection in force was a type of sexual selection in which females preferred only males of a certain genotype with a high probability. In an experiment in which only this type of selection was in force, it was shown that this sexual component of natural selection was sufficient for a mutant allele that produced the preferred genotype to drive this genotype to predominance in both sexes in a population that evolved from a small founder population in which the sexually preferred genotype was not present but had arisen from a mutation during the evolution of the two-sex population.

An obvious observation that may be made on any human population and many animal populations is that at any time the population consists of overlapping generations corresponding to age cohorts in both females and males. It is thought by many that this overlapping of generation provided milieu for the passing on of culture from older to younger individuals and the evolution of high human cognitive abilities when compared to other animals. Chapter 12 is devoted to the development of a self-regulating two-sex stochastic population process that accommodates the evolution of an age structured population along with the genetics underlying such populations. From the vantage point of branching process, the formulation in chapter 12 is an algorithmic extension of what is known as the general branching process. Among the components of selection included in this age structured formulation are the parametrized risk functions of death by each age group in both the female and male populations. As of the complexity of this formulation, attention was confined to the embedded deterministic model for all computer experiments reported in this chapter, but as time passes attention will be devoted to problems centered around the development of algorithms to compute Monte Carlo realizations of the age structured process.

An ultimate goal of the research on the stochastic models of evolution presented in this book is to include the evolution of a genome at the molecular level and how processes of natural and artificial selected may have affected the structure of genomes. Before undertaking research to include models of genomes in stochastic models of evolution, it seems necessary to develop some acquaintance with recent literature on the concept of a gene. Accordingly, chapter 13 is devoted a review of biological literature on the history of the concept of gene and a peek at one of the most recent definitions of a gene. Basically, a region of DNA that codes for either proteins or $RNAs$ are frequently composed of introns and exons, and during the process of transcription and subsequent processing of the products of transcription, exons are sliced together in alternative ways to produce proteins that are used in various stages of development of an individual. But, the process of transcription is not an end in itself but is evidently controlled by regulatory regions of genomic DNA where other chemical structures, that may be coded for by other genes, bind and turn the transcription process of a given gene on and off. Thus, in order to get a grip as to how mutation may affect a gene, one must take into account not only the coding region of DNA but also those regulatory regions of DNA that turn coding regions off and on. The union of such regions of genomic DNA may involve thousands or even millions of bases as described in actual biological examples of genes in the closing sections of chapter 13.

Chapter 14, which is a small book within a book, is organized around three themes. One theme concerns a review of recent literature on developing computer models designed to simulate the evolution of model genomes that may consists of as many of a million or more base pairs in the presence of such forces of evolution as mutation and selection. Among those models reviewed, the accounts as to how mutation, selection and genetic recombination were incorporated into their formulations were not given is sufficient detail to suit members of the community of mathematical sciences, who may wish to replicate their reported results.

A second theme was a review of new methods for detecting signals of evolution in human haplotype data. This very recent research involved a model of genomic evolution that ran backwards in time to some founder population that existed about 40,000 or so years ago. Moreover, distributions of scores, representing measure of selection, which were estimated from simulated genomic data produced by the backwards in time simulation model, were applied in existing human haplotype data to detect signals of selection in relatively small regions of a genome that heretofore had not been possible to detected.

A third theme involved a phylogenetic study of several mammalian species designed to detect signals of natural or artificial selection in protein coding genes in several mammalian species whose genomes had been sequenced. The methods used in this research involved, among other things, a evolutionary model of three letter codons related to those discussed in chapter 6. Briefly, maximum likelihood estimates of parameters in these models were used to construct indicators of signals of selection in protein coding regions of the species under consideration.

Finally, the fourth theme of this chapter was centered around the problem of constructing stochastic models to simulate various types of mutations, that occur at the genomic level, as well as genetic recombination that included crossing over and gene conversion with respect to two makers during a phase of meiosis in diploid population when chromatids are arranged along a spindle of a dividing cell. Similarly, the various types of mutations, whose mathematical structure was thoroughly documented, were assumed to occur during meiosis and in that phase when the DNA content of the cell is doubled by a process that may involve that errors during the DNA copying process. The types of mutation included in this chapter were nucleotide substitution, deletions and insertions, copy number changes in finite segments of DNA and inversions of segments of DNA within a chromosome. Both the models of genetic recombination and mutations are documented in sufficient detail so a read can discern their details and either accept, modify or reject a proposed formulation. This process acceptance, modification or rejection in a time honored way to progress in the development of models that eventually find acceptance in a given field of science.

The book ends with chapter 15, which is devoted to a suggested agenda for continuing the research projects proposed in the preceding chapters and a short review of books and material from other media that were sources of inspiration while writing and thinking about the contents of this book.

This page intentionally left blank

Acknowledgments

A number of people have been helpful in the writing and assembling the material used in this book. Of particular note is Towfique Raj, who as a undergraduate student at Drexel University requested a course from the senior author in Mathematical Genetics in the winter quarter of 2004, and it was his enthusiasm for the subject that awakened dormant interest in Genetics that eventually led to the writing of this book. Presently, Towfique is on a Postdoctoral Research Fellow, Harvard Medical School, Brigham Women's Hospital, Boston, MA 02115, and was very influential in assisting the authors in obtaining references and graphs for some of the material presented in chapter 14. Conversations with Warren Ewens of the University of Pennsylvania were also helpful in that he suggested that more work needed to be done on the modelling of selection and he has also read drafts of some of the chapters and offered suggestions for improvements.

Several contacts with working geneticists and others interested in evolution were also made at a symposium on evolution that was held in May and early June of 2009 at Cold Spring Harbor Laboratory on Long Island, New York in connection with the celebration of the 200-th anniversary of Charles Darwin's birth. Among the contacts at this symposium was Mike Levine of the University of California, Berkeley, who called attention to the mouse sonic hedgehog gene that was regulated by an enhancer 1 megabase away and would thus be a good biological example in thinking about computer models of genes as indicated in chapter 13. David Shaw, Mouse Genome Informatics, The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609, was also helpful also obtaining more information on this gene. David Gasser and Douglas Epstein, Geneticists at the University of Pennsylvania, were also very helpful in obtaining further information of this gene as well as information on an ortholog of this gene in humans, which has been implicated in holoprosencephaly.

Thanks are also due Peter Olofsson, Department of Mathematics, Trinity University, who read and offered suggestions for improvements of chapter 9, which were incorporated into the present version of this chapter. Margaret Dominy, Science Information Services, Drexel University Library, was also very helpful in providing copies of papers in older volumes of scientific journals that had not been digitized and were thus not available on the internet and also copies of papers in more recent publications on the internet. A professional proof reader and friend, the late Robert (Bob) Grant, also proof read several chapter of the book, which resulted in the correction of many minor errors. Another friend, Scott Frizen, read chapter 1 with close attention to details and offered suggestions that were incorporated into the present version of chapter 1. It is also fitting to acknowledge that the internet search engine, Google, played a significant role in finding many of the papers and other material cited in this book. Special thanks are also due to MATHWORKS, who permitted the use MATLAB free of charge in producing most of the graphs in the book. And finally, warm thanks to George Pearson and John MacKendrick of MacKichan Software, Inc., who provided invaluable and timely assistance in using the technical word processor, Scientific Workplace, which was used to write and produce the book.

Contents

Chapter 1

An Introduction to Mathematical Probability with Applications in Mendelian Genetics

1.1 Introduction

Genetics is a major and basic field of biology with a vast literature. Indeed this literature is so vast that it is doubtful if any person has a complete grasp of the available material on genetics in numerous books and on the world wide web. A current and excellent introduction to genetics is contained in the text book, Principles of Genetics, by Snustad and Simmons (2006), and it is suggested that if a reader encounters a term that is not familiar, either this book, a similar book or the world wide web be consulted. In this chapter, however, it will be assumed that a reader has a grasp of such terms from Mendelian genetics as gene, genotype, chromosomes and phenotype, and in subsequent chapters other terms will be introduced as attempts are made to deal with problems in genetics that arise at the molecular level.

Genetics deals with biological processes that are not deterministic. That is, the present state of a process does not fully determine the next state. For example, suppose a couple marries and intends to have children, but, given this state of matrimony, it is impossible to predict with certainty how many children they may have or the distribution of boys and girls in their family. There are also many examples in medicine that are characterized by uncertainty. A doctor may, for example, administer the same treatment to patients with a disease of the same phenotype, but the patients will react differently to the same treatment.

Processes that are characterized by uncertainty or randomness are called stochastic, a term that is derived from the Greek word which means to aim at or characterized by randomness. From the mathematical point of view, stochastic processes are rooted in concepts and procedures of probability theory, a field of mathematics that deals with the concept of uncertainty.

The purpose of this chapter is to introduce a reader to basic probability theory and some of its famous distributions that have been and will continue to be applied in the study of genetic and evolutionary phenomena. Should a reader find some of the terminology unfamiliar, it is suggested that the references cited or information on the world wide web be consulted. As the terms in this chapter closely follow terminology that is used widely in genetics and probability, a web search may be expedited by entering a key word into a search engine.

1.2 Mathematical Probability in Mendelian Genetics

Mathematical probability is concerned with the analysis of the intuitive notion of chance or randomness. Anybody with some acquaintance with the laws of genetics encounters the notion of chance or randomness in the interpretation of observed phenomena. For example, let A denote an autosomal gene dominant to its allele a and suppose a male of genotype Aa is mated to a female of genotype Aa.

A mating of the type $Aa \otimes Aa$ is capable of producing three types of offspring; namely AA, Aa, and aa. Before a particular offspring is born or comes into existence, we cannot predict his genotype with certainty, although experience has shown that in large numbers of offspring from a mating of the type $Aa \otimes Aa$ 1/4 are of genotype AA, 1/2 are of genotype Aa, and 1/4 are of genotype aa. We express these empirical facts by saying the probability of getting an offspring of genotype AA, Aa, or aa from a mating of the type $Aa \otimes Aa$ is 1/4, 1/2, and 1/4, respectively.

Closer analysis of the mating $Aa \otimes Aa$ shows that there are four basic events in the production of offspring with genotypes AA, Aa, and aa. Offspring of genotypes AA or aa arise if, and only if, both the male and the female contribute like alleles, and an offspring of phenotype Aa arises only when both the male and female contribute different alleles. The situation may be represented diagrammatically by the table 1.2.1.

Table 1.2.1 Gametes and Offspring of Mating Aa ⊗ Aa

$\cdot$	A	a
A	AA	Aa
a	aA	aa

The symbols in the second and third positions in the first row of the table represent the gametes A and a contributed by the female. Similarly, the symbols in the first column of the table represent the gametes contributed by the male. The four basic events mentioned above are represented in the four remaining cells of the table. For example, the genotype Aa represents a case in which allele A was contributed by the male and allele a was contributed by the female. The genotypes Aa and aA have the same phenotype.

From the nature of meiosis, the complex cell division leading to the formation of the sex cells, we would expect both the male and female, in the absence of disturbing influences, to produce the gametes A and a in equal proportions or with probabilities equal to $1/2$. These empirical facts would in turn lead us to assign a probability of $1/4$ to each genotype in the table. The probabilistic model which has just been described is a concrete genetic realization of a mathematical structure called a probability space. The demonstration of this statement will be deferred until we have a better understanding of what is meant by a probability space.

When dealing with a probability space we are concerned with three things. Firstly, the set Ω of all logically possibly outcomes of a conceptual experiment, secondly, a collection $\mathcal{A}$ of subsets in Ω called events, and thirdly a set function P defined on $\mathcal{A}$, i.e., for every event A in $\mathcal{A}$ we assign a number $P\,[A]$ called the probability of A. But, before we can give a precise definition of a probability space, it will be necessary to study the algebra of sets or events as they are called in probability theory. With the passage of time, it is becoming increasingly clear that the algebra of sets will play a helpful role in the development of concepts to deal with complex situations that arise in genetics and evolution.

An event is simply some subset of the sample space Ω, which for the moment will be assumed to contain a finite number N of elements denoted by ω_i for $i = 1, 2, 3, \ldots, N$. For example, if Ω is the set $(\omega_1, \omega_2, \omega_3, \omega_4)$, then the sets (ω_1, ω_2) and (ω_2, ω_4) as well as Ω itself are events. If an event contains a single point ω, then we shall write (ω), and arbitrary events will be denoted by the letters A, B, C, D, $E, \ldots$ with or without subscripts. We say an event A occurs if, and only if, a point ω belonging to A is observed, and the statement "ω belongs to A" shall be written as $\omega \in A$.

To each event A there corresponds the contrary event not A, i.e., the event that A does not occur. The event not A is called the complement of A and will be denoted by A^c. For example, if Ω is the set given in the previous paragraph and $A = (\omega_1, \omega_2)$, then $A^c = (\omega_3, \omega_4)$.

If an event A implies the occurrence of the event B, we shall write $A \subset B$, i.e., A is a set contained in B and if A occurs then B necessarily occurs. If at the same time the event B implies the occurrence of the event A, then we say the events A and B are equivalent and write $A = B$. As an example of the notion of an event being contained in another, consider the events $A = (\omega_2, \omega_3)$ and $B = (\omega_2, \omega_3, \omega_4)$.

The notion of equality of events is, of course, useful only in the case the two events under consideration contain large numbers of points so that the question of equality cannot be decided by an easy enumeration. An equivalent way of writing the relation $A \subset B$ is $B \supset A$ which is read B contains A.

Events may be combined into new events by means of operations expressed by the words "and", "or", and "not". The event A "and" B is an event which occurs if, and only if, both events A and B occur. For the event A "and" B we shall write $A \cap B$ which reads A intersection B or the intersection of A and B. Frequently, for the sake of brevity, we shall drop the symbol $\cap$ and write AB for A intersection B or the intersection of A and B. As an example of the notion of intersection of sets, if $A = (\omega_2, \omega_3)$ and $B = (\omega_3, \omega_4)$, then their intersection is $AB = (\omega_3)$. In general, for any $\omega \in \Omega, \omega \in A \cap B$ if, and only if, $\omega \in A$ and $\omega \in B$.

The idea of the impossible or null event, which will be denoted by φ, plays an important role in the algebra of events or sets. For example, if the event AB cannot occur, then we say the events A and B are disjoint or mutually exclusive. This notion may be symbolized by setting $AB = \varphi$, i.e., the occurrence of A and B is an impossible event. Observe that for any event A, it follows that $A \cap A^c = \varphi$.

The event A "or" B occurs if, and only if, at least one of the events A, B occurs. We shall write $A \cup B$ for this event which reads A union B. If the events A and B are disjoint so that $AB = \varphi$, we shall write $A + B$ for A union B. As an example of the notion of union of events let $A = (\omega_2, \omega_3)$ and $B = (\omega_3, \omega_4)$. Then, $A \cup B = (\omega_2, \omega_3, \omega_4)$.

More generally, if we are considering some finite collection of events say $A_1, A_2, \ldots, A_n$ for $n < N$, then we shall write

$$\bigcap_{i=1}^{n} A_i \tag{1.2.1}$$

for the intersection of these events or the event that all of the n events occur. For the event at least one of these events occur, we shall write

$$\bigcup_{i=1}^{n} A_i \tag{1.2.2}$$

for the union of the n events. We say this collection of events is disjoint if $A_i \cap A_j = \varphi$ for $i \neq j$. For the case of disjoint events, we may write

$$\sum_{i=1}^{n} A_i \tag{1.2.3}$$

for their union or the event that at least one of the events occurs. It sometimes happens that we wish to consider the intersection of two unions of events, for example, the intersection of $A_1 \cup A_2$ and $A_5 \cup A_4$.

For an intersection of this kind, we shall write $(A_1 \cup A_2) \cap (A_5 \cup A_4)$ or equivalently $(A_1 \cup A_2)\,(A_5 \cup A_4)$. From the definitions which have been given so far it follows that if $A \subset B$, then $B^c \subset A^c$. Furthermore, for any event or set A under consideration $A \subset \Omega$, and $\Omega^c = \varphi$ so that $\varphi^c = \Omega$.

The algebra of events or sets is governed by a few simple laws. A list of the more frequently encountered laws follows. All of the laws listed below could be proved as theorems, but we shall not take the time and space to carry out proofs. If the reader has not encountered these laws before, it would be profitable to illustrate them by examples.

Commutative Laws

$$A \cap B = B \cap A$$
$$A \cup B = B \cup A$$

Associative Laws

$$(AB)C = A(BC)$$
$$(A \cup B) \cup C = A \cup (B \cup C)$$

Distributive Laws

$$A(B \cup C) = AB \cup AC$$
$$A \cup BC = (A \cup B)(A \cup C)$$

Identity Laws

$$A \cup \varphi = A$$
$$A\Omega = A$$

Idempotent Laws

$$AA = A$$
$$A \cup A = A$$

De Morgan's Laws

$$(AB)^c = A^c \cup B^c$$
$$(A \cup B)^c = A^c B^c$$

Upon close observation one can perceive an important principle known as the Duality Law. Every valid relation amongst sets obtained by taking complements, unions, and intersections is transformed into a valid relation if the symbols "=" and "c" remain unchanged and the symbols $\cap$, $\subset$, and φ are interchanged with the symbols $\cup$, $\supset$, and Ω, respectively.

It will be noted that each of the above laws occur in pairs and that a given statement in a pair is the dual of the other. Sometimes it is possible to derive less obvious laws from more obvious laws by an application of the Duality Law. For example, the first distributive law $A(B\cup C) = AB\cup AC$ is obvious enough; its dual $A\cup BC = (A\cup B)(A\cup C)$ is more subtle, however.

Similarly, for every event A the relation $A \subset \Omega$ is true and obvious. Its dual statement $A \supset \varphi$ or equivalently $\varphi \subset A$ is much more subtle. A thorough justification of the laws of the algebra of sets depends on the discussion of points in formal logic, but we shall not pursue these matters here.

We are now ready for a formal definition of a mathematical structure which plays a basic role in mathematical genetics; namely a finite probability space. Let Ω be the finite set of all logical outcomes of a conceptual experiment. Points in Ω will be denoted by ω. A finite probability space is the triple $(\Omega, \mathcal{A}, P)$ satisfying the following conditions.

I. The collection $\mathcal{A}$ of events in Ω is an algebra, i.e., a class of sets closed under all finite set operations. More precisely, if $A \in \mathcal{A}$ (read A is a member of $\mathcal{A}$}, then $A^c \in \mathcal{A}$ and if $A_i \in \mathcal{A}$ for $i = 1, 2, \ldots, n$, then

$$\bigcap_{i=1}^{n} A_i$$

and

$$\bigcup_{i=1}^{n} A_i$$

belong to $\mathcal{A}$.

II. P is a set function defined on the algebra $\mathcal{A}$, which satisfies the following conditions. $P[\Omega] = 1$ (normed property) For every event $A \in \mathcal{A}$, $P[A] \geq 0$ (non-negative property). For every disjoint class of sets $A_1, A_2, \ldots,\ A_n$ in the algebra $\mathcal{A}$.

$$P\left[\sum_{i=1}^{n} A_i\right] = \sum_{i=1}^{n} P[A_i],$$

which is called the finitely additive property.

The function $P[\cdot]$ on $\mathcal{A}$ is frequently called a probability measure. A number of other properties are easily deducible from properties ***I*** and ***II***, which define a finite probability space. Some of the most frequently encountered properties are enumerated below.

Let A be any event in $\mathcal{A}$. Then, $A^c \in \mathcal{A}$. Furthermore, $A + A^c = \Omega \in \mathcal{A}$ so that $\Omega \in \mathcal{A}$, Consequently, $\Omega^c = \varphi \in \mathcal{A}$.

It also follows from ***I*** and ***II*** that $P[\varphi] = 0$. To see this observe that for any $A \in \mathcal{A}$, $A + \varphi = A$. From the additivity property of P, it follows that $P[A] + P[\varphi] = P[A]$, which implies that $P[\varphi] = 0$

If $A \subset B$, then $B = A + A^cB$. and by the additivity property of P, it follows that$P[B] = P[A] + P[A^cB]$ or

$P[A] = P[B] - P[A^cB] \leq P[B]$, because $P[A^cB] \geq 0$. Since $A \subset \Omega$ for every $A \in \mathcal{A}$, it follows that $P[A] \leq P[\Omega] = 1$ for all $A \in \mathcal{A}$.

Any arbitrary finite union of sets in $\mathcal{A}$

$$\bigcup_{i=1}^{n} A_i \in \mathcal{A} \tag{1.2.4}$$

may be written as a disjoint union. To verify this statement, consider the case $n = 3$. Then,

$$A_1 \cup A_2 \cup A_3 = A_1 + A_2A_1^c + A_3A_1^cA_2^c. \tag{1.2.5}$$

It is easy to see that this scheme may be generalized to any $n > 3$. Therefore,

$$P[A_1 \cup A_2 \cup A_3] \leq P[A_1] + P[A_2] + P[A_3]. \tag{1.2.6}$$

In general, for any arbitrary finite union of sets in $\mathcal{A}$, it follows that

$$P\left[\bigcup_{i=1}^{n} A_i\right] \leq \sum_{i=1}^{n} P[A_i] \tag{1.2.7}$$

This inequality is known as Boole's inequality.

Finally, for any $A \in \mathcal{A}$, $P[A^c] = 1 - P[A]$. To see this note that $A + A^c = \Omega$ so that $P[A] + P[A^c] = 1$, which implies that $P[A^c] = 1 - P[A]$.

In the next section, several concrete examples of finite probability spaces will be given that are useful in genetics.

1.3 Examples of Finite Probability Spaces

In this section some concrete examples of the finite probability spaces will be given. As these examples will illustrate, the concept of a finite probability space may take many forms.

Example 1.3.1: An Equal Frequency Model

Historically, this is the model which gave rise to the conditions that were used to define a finite probability space in the previous section. Let Ω be any set containing the finite number of points $(\omega_1, \omega_2, \ldots, \omega_N)$, where N is a positive integer. The algebra $\mathcal{A}$ of sets in Ω may be chosen as follows. For the case $N = 3$, consider the class $\mathbb{C}$ of all subsets of Ω.

A straight-forward enumeration of this class leads to the conclusion that

$$\mathbb{C} = \{\varphi, (\omega_1), (\omega_2), (\omega_3), (\omega_1, \omega_2), (\omega_1, \omega_3), (\omega_2, \omega_3), \Omega\}. \tag{1.3.1}$$

Observe that this class contains $8 = 2^3$ sets and it is closed under all finite set operations involving unions, intersections and complements. Therefore, the algebra $\mathcal{A}$ may be chosen as $\mathcal{A} = \mathbb{C}$. Subsequently, it will be shown that for any integer $N \geq 1$, the class $\mathbb{C}$ of all subsets of Ω contains 2^N sets and it suffices to choose $\mathcal{A} = \mathbb{C}$.

In general, to define a probability P on $\mathcal{A}$ let

$$P[\omega] = \frac{1}{N} \tag{1.3.2}$$

for all $\omega \in \Omega$. Then, for every set $A \in \mathcal{A}$, let

$$P[A] = \sum_{\omega \in A} P(\omega). \tag{1.3.3}$$

By definition, if $A = \varphi$, then $P[A] = 0$. Given these definitions, it is easy to verify that the triple $(\Omega, \mathcal{A}, P)$ is a finite probability space.

Comment: The probabilistic model discussed in the previous section, characterizing the offspring obtainable from the mating $Aa \times Aa$, is a concrete example of an equal frequency model considered above. Specifically, Ω is the set of genotypes in the body of Table 1.1.1 and a probability of 1/4 is assigned to each of the four genotypes.

Example 1.3.2: Partitions of an Abstract Set Ω

Sometimes when one wishes to construct a finite probability space but it is difficult to conceptualize the set Ω of all possible events, it is often useful to consider a finite collection of events such that it forms a partition of Ω, which is easy to conceptualize. For example, let A_i for $i = 1, 2, \ldots, N$ be a disjoint collection of events in Ω such that their disjoint union is Ω.

Symbolically, this condition is expressed as

$$\sum_{i=1}^{N} A_i = \Omega. \tag{1.3.4}$$

Such disjoint classes of events are called partitions of Ω.

Starting with such a partition of Ω, one may construct a finite probability space according to the following scheme. Let $\mathcal{A}$ be the collection of all finite sums (disjoint unions) of sets in the partition $(A_1, A_2, \ldots, A_N)$.

The sum containing no sets will be, by definition, the impossible event. It is easy to verify the collection $\mathcal{A}$ so defined is an algebra. If we assign the probability $P[A_i] = p_i$, where $p_i \geq 0$ and

$$\sum_{i=1}^{N} p_i = 1, \tag{1.3.5}$$

to each event in the partition $(A_1, A_2, \ldots, A_N)$, then a probability for each event $E \in \mathcal{A}$ may be assigned as follows. Let $A_{i_1}, A_{i_s}, \ldots, A_{i_k}$ denote the disjoint events in E. Then,

$$P\left[E\right] = \sum_{\nu=1}^{k} P\left[A_{i_\nu}\right] \tag{1.3.6}$$

for every $E \in \mathcal{A}$. It is clear that a triple $(\Omega, \mathcal{A}, P)$ so defined is a finite probability space.

Comment: If Ω is thought of as all those events that occur at a molecular and cellular levels that lead to the production of offspring in a mating of the form $Aa \times A_a$, then the equal frequency model described above is a special case of the more general model described in this example. At the molecular and cellular level, Ω may be thought of as a black box.

Example 1.3.3: A Deterministic Case

In the deterministic case we are concerned with a situation in which an event always occurs.

This example is actually a special case of the situation discussed in example 1.3.2, but it will be treated separately in order to show a deterministic situation may be included within the framework of probability theory.

Let Ω denote an event that always occurs. In terms of probability, this condition is expressed by putting $P\left[\Omega\right] = 1$ and $P\left[\Omega^c\right] = P\left[\varphi\right] = 0$. Then the triple $(\Omega, \mathcal{A}, P)$, where $\mathcal{A} = (\varphi, \Omega)$, is a finite probability space.

Example 1.3.4: Inheritance of Eye Color and Sex

For the sake of emphasis, another simple example from genetics will be given. Let B be a dominant autosomal gene for brown eyes, let b be its recessive allele for blue eyes, and let X and Y stand for the sex chromosomes. We shall suppose the human population is under discussion so that an individual carrying two X chromosomes is female and an individual carrying a X and a Y chromosome is male. The reader should not be misled into believing that the inheritance of eye color in man is as simple as is stated here. In reality it is much more complex, although the model we are considering may apply in certain families.

Our aim is to construct a probability space characterizing the production of offspring from the mating $BbXX \otimes BbXY$. The set of logical outcomes from this mating may be represented by the table 1.3.1.

Table 1.3.1 Gametes and Offspring from Mating BbXX ⊗ BbXY

$\cdot$	$\boldsymbol{BX}$	$\boldsymbol{BY}$	$\boldsymbol{bX}$	$\boldsymbol{bY}$
BX	$BBXX = \omega_1$	$BBXY = \omega_2$	$BbXX = \omega_3$	$BbXY = \omega_4$
bX	$BbXX = \omega_5$	$BbXY = \omega_6$	$bbXX = \omega_7$	$bbXY = \omega_8$

The two symbols in column one of the table represent the two gametes produced by a female of genotype $BbXX$. Similarly, the four symbols in the first row of the table represent the types of gametes produced by a male of genotype $BbXY$. The eight basic events, representing the various genotypes that may be produced by the mating $BbXX \otimes BbXY$, are given in the body of the table. Since the allelic pair (B, b) is autosomal, we would expect both the male and female to produce the possible types of gametes in equal proportions, which in turn would lead to assign a probability of $1/8$ to each basic event. We thus have another genetic realization of the equal frequency model of example 1.3.1. The probability space characterizing the situation is, therefore, the space discussed in this example with $N = 8$.

After constructing a probability space it is always of interest to ask for the probabilities of events which may be described in terms of English sentences or statements. Some examples of such questions are given below.

1. What is the probability of the event, "a blue eyed girl"? We see this event is the basic event (ω_7). Hence, $P[\omega_7] = 1/8$.

2. What is the probability of the event "a brown eyed boy"? The event "a brown eyed boy" is the union of the basic events (ω_2), (ω_4) and (ω_6). Hence, $P\left[(\omega_2, \omega_4, \omega_6)\right] = 3/8$.

3. What is the probability of the event "the offspring is a boy, he is heterozygous for brown eyes, and obtained his gene for brown eyes from his mother"? We see this event corresponds to the basic event (ω_4) so that $P[(\omega_4)] = 1/8$.

It is interesting to consider a relationship between the set of all statements describing the events of a conceptual experiment and the class of events in the algebra $\mathcal{A}$. Unfortunately, there is no one-to-one correspondence between statements and events, for many statements may correspond to the same event. For example, in the situation just discussed and in the realm of ordinary human experience the statements "The offspring is a girl with three arms" and "The offspring is a boy with four arms" both correspond to the impossible event φ. In working practical problems in probability, our most difficult task is often that of deciding which event in $\mathcal{A}$ corresponds to a given statement.

So far we have constructed probability spaces which provide answers to only very simple questions. We would also like to be able to give answers to questions of the following kind. In families of size four from a mating of the type $BbXX \otimes BbXY$ what is the probability of the events (i) "four brown eyed boys" and (ii) "three blue eyed girls and a brown eyed boy". To answer questions of this kind more elaborate probability spaces will be required. To construct such probability spaces in a simple and elegant way, it will be necessary to discuss some concepts from elementary combinatorial analysis.

1.4 Elementary Combinatorial Analysis

Throughout the following discussion we shall consider a set Ω containing $N \geq 1$ elements, $\omega_1, \omega_2, \ldots, \omega_N$. A permutation of a set is an ordering of its elements. If $N = 2$ for example, then the set has two permutations; namely $\omega_1\omega_2$ and $\omega_2\omega_1$. The permutation, $\omega_1\omega_2 \ldots \omega_N$ is called the natural ordering. In general, if a set Ω contains $N \geq 1$ elements, then it can be ordered in

$$N! = N \times (N-1) \times (N-2) \times \ldots \times 2 \times 1 \tag{1.4.1}$$

ways. The symbols $N!$ is read N factorial.

To prove this statement, observe that the first element of a permutation may be chosen in N ways, the second element in $(N-1)$ ways, the third element in $(N-2)$ ways and so on. The number of permutations of a set is the product of the number of ways each element in the permutation may be chosen. It follows, therefore, that the number of permutations of a set containing Ω containing $N \geq 1$ elements is $N!$.

If a set Ω has N elements, then $k \leq N$ elements of this set can be ordered in

$$_{N}P_{k} = \frac{N!}{(N-k)!)} \tag{1.4.2}$$

ways. To prove this statement, note that the first element may be chosen in N ways, the second in $(N-1)$ ways and so on down to element k which can be chosen in $(N-(k-1)) = N-k+1$ ways. Therefore,

$$_{N}P_{k} = N \times (N-1) \times \ldots \times (N-k+1) = \frac{N!}{(N-k)!)} \tag{1.4.3}$$

A combination is any non-ordered subset of a set containing a given number $k \geq 0$ of elements. By definition, the combination containing $k = 0$ elements is the empty set. For k such that $0 \leq k \leq N$, let the symbol

$$\binom{N}{k} \tag{1.4.4}$$

denote the number of combinations of size k of a set containing N elements. Then,

$$\binom{N}{k} = \frac{N!}{k!\,(N-k)!)}. \tag{1.4.5}$$

For the case $k = 0$, k factorial will be defined as $0! = 1$. With this definition, it follows that

$$\binom{N}{0} = 1, \tag{1.4.6}$$

indicating that there is only one empty set in the class $\mathbb{C}$ of all subsets of a set Ω containing N elements.

To prove the validity of formula (1.4.5) , observe that

$$k!\binom{N}{k} = \frac{N!}{(N-k)!)}, \tag{1.4.7}$$

which is equivalent to (1.4.5) . If a set Ω contains N elements, let M denote the number of sets in the class $\mathbb{C}$ of all subsets of Ω. Then, it follows that M is given by the equation

$$M = \sum_{k=0}^{N} \binom{N}{k}. \tag{1.4.8}$$

It is of interest to investigate whether is sum can be reduced to a simple formula for M as a function of N.

The numbers $\binom{N}{k}$ are also known as the binomial coefficient because they appear in the binomial expansion $(a+b)^{N}$, where a and b can be any real numbers. The symbol $(a+b)^N$ means that the sum $a+b$ is multiplied by itself $N \geq 1$ times. Symbolically, this product may be represented by the N-fold product.

$$(a+b) \times (a+b) \times \ldots \times (a+b)\,. \tag{1.4.9}$$

A typical product term in this expansion containing N symbols will contain the letter "a" k times and the letter "b" $N-k$ times, which reduces to the formula $a^k b^{N-k}$. However, the number of ways one can choose k places of the symbol a is the number of subsets of size k that may be chosen from a set containing N elements. In the discussion above, it was shown that this number was $\binom{N}{k}$. Therefore,

$$(a+b)^N = \sum_{k=0}^{N} \binom{N}{k} a^k b^{N-k}. \tag{1.4.10}$$

This famous formula is known as the binomial theorem. In particular, if $a = b = 1$, then this formula becomes

$$2^N = \sum_{k=0}^{N} \binom{N}{k} = M, \tag{1.4.11}$$

which is the number of sets in the class $\mathbb{C}$ of all subsets of a set Ω containing N elements.

The number of sets in the class $\mathbb{C}$ can be very large even for moderate values of N. For example, for $N = 30$

$$M = 2^{30} = 1{,}073,\ 741,\ 824. \tag{1.4.12}$$

When M is a number of this size, over one billion, it would be a very challenging problem to write and execute computer code to enumerate all the sets in the class $\mathbb{C}$.

To complete our discussion of combinatorial analysis, it will be necessary to consider partitions of a set Ω containing $N \geq 1$ elements. A collection of disjoint sets $A_1, A_2, \ldots, A_r$ in Ω for $r \geq 3$ such that each set contains $k_1, k_2, \ldots, k_r$ elements, respectively, is said to be a partition of Ω if

$$k_1 + k_2 + \cdots + k_r = N \tag{1.4.13}$$

and their union is Ω. In symbols,

$$\sum_{\nu=1}^{r} A_\nu = \Omega. \tag{1.4.14}$$

For some ν, k_ν may be zero so that some of the sets in a partition may be empty.

For a set Ω containing $N \geq 1$ elements, the number of partitions of the type just described is given by the formula

$$\frac{N!}{k_1! \times k_2! \times \ldots \times k_r!}. \tag{1.4.15}$$

To prove the validity of this formula, observe that the k_1 elements in the set $A_!$ may be chosen in $\binom{N}{k_1}$ ways. Then, after the elements of A_1 have been chosen, the elements of the set A_2 may be chosen in $\binom{N-k_1}{k_2}$ ways. By continuing in this way, it can be seen that the elements in the set A_r may be chosen in $\binom{N-k_1-k_k-\ldots k_{r-1}}{k_r} = 1$ way, Therefore, the number in question is

$$\begin{aligned}&\binom{N}{k_1} \times \binom{N-k_1}{k_2} \times \ldots \times \binom{N-k_1-k_k-\ldots k_{r-1}}{k_r} \\ &= \frac{N!}{k_1! \times k_2! \times \ldots \times k_r!}.\end{aligned} \tag{1.4.16}$$

These numbers are also known as the multinomial coefficients, because they appear in the multinomial theorem.

Let a_ν for $\nu = 1, 2, \ldots, r$ for a set of real numbers. Then, the multinomial theorem is the statement or equation

$$(a_1 + a_2 + \cdots + a_r)^N = \sum \frac{N!}{k_1! \times k_2! \times \ldots \times k_r!} a_1^{k_1} a_2^{k_2} \times \ldots \times a_r^{k_r}, \tag{1.4.17}$$

where the sum extends over all solutions of the equation

$$k_1 + k_2 + \cdots + k_r = N \tag{1.4.18}$$

such that $k_\nu \geq 0$ is a non-negative integer for $\nu = 1, 2, \ldots, r$. The proof of this theorem is very similar to that for the binomial theorem and will, therefore, be omitted.

Given equation (1.4.18) a problem that naturally arises is that of finding the number of solutions of this equations. For any fixed positive integers r and N, the number of solutions to this equation is given by the formula

$$\binom{N+r-1}{N} = \binom{N+r-1}{r-1}. \tag{1.4.19}$$

Equivalently, this formula also gives the number of terms in the expansion of $(a_1 + a_2 + \cdots + a_r)^N$.

To demonstrate the validity of this formula, consider the case $r = 5$ and $N = 6$. Then, imagine an experiment in which 6 balls are dropped at random into 5 cells. The five cells will be represented by six bars $|$ and the six balls by stars $*$. Then, the configuration $| * | ** || ** | * |$ would represent a solution of equation (1.4.18) such that $k_1 = 1, k_2 = 2, k_3 = 0, k_4 = 2$ and $k_5 = 1$. In general, r cells may be represented by $r + 1$ bars and N balls may be represented by N stars. Each configuration of this kind starts and ends with a bar, but the remaining $N + r - 1$ symbols may be arranged in any way. It thus becomes clear that the number of solutions of equation (1.4.18) correspond to the number of ways one can choose N positions to put the balls from the $N + r - 1$ positions available, which is the number on the left in (1.4.19).

1.5 The Binomial Distribution

Having discussed some points in elementary combinatorial analysis, we are now ready to consider an important distribution in mathematical genetics and indeed in probability theory in general; namely the binomial distribution. In this distribution the notion of independence plays an important role so, we begin by giving a formal definition of this concept.

If $(\Omega, \mathcal{A}, P)$ is any probability space, then events A_1 and A_2 in $\mathcal{A}$ are independent if, and only if, $P[A_1 A_2] = P[A_1]P[A_2]$. More generally, if $A_1, \ldots.\ A_N$ are some finite number $N \geq 2$ of events in $\mathcal{A}$, then these events are independent if, and only if, for every integer r such that $2 \leq r \leq N$ and every collection $A_{k_1}, \ldots, A_{kr}$, of events from the N events, we have

$$P\left[A_{k_1} A_{k_2} \ldots A_{k_r}\right] = P\left[A_{k_1}\right] \times P\left[A_{k_2}\right] \times \ldots \times P\left[A_{k_r}\right]. \tag{1.5.1}$$

It should be noted that this definition of independence imposes $2^N - N - 1$ conditions on the probabilities associated with the N events $A_1, \ldots, A_N$. The reason for formulating the definition of independence in this way is because it is possible, for example, to have the multiplication rule hold for all events taken two at a time but not for all events taken three at a time and so on.

We have, of course, already tacitly encountered the notion of independence in the construction of probability spaces characterizing the production of offspring from a mating. Upon inspection of previous examples, see Table 1.3.1, it will be observed that we have assumed the gametes of the parents combine independently to produce the genotypes of the offspring. The statement below gives an interesting consequence of the definition of independence for the case of two events.

If the events (A_1, A_2) are independent, then so are the pairs of events (A_1^c, A_2), (A_1, A_2^c), and (A_1^c, A_2^c). To demonstrate the validity of this statement by an application of the identity and distributive laws, it follows that

$$A_2 = A_2\Omega = A_2\left(A_1 + A_1^c\right) = A_1A_2 + A_1^cA_2. \tag{1.5.2}$$

And, by additivity of probabilities, it follows that

$$P\left[A_2\right] = P\left[A_1A_2\right] + P\left[A_1^cA_2\right]. \tag{1.5.3}$$

However, under the hypothesis of independence of events (A_1, A_2), it follows that $P[A_1A_2] = P[A_1]P[A_2]$. Using this condition and solving for $P\left[A_1^cA_2\right]$, we find that

$$P\left[A_1^cA_2\right] = P\left[A_2\right] - P[A_1]P[A_2] = P\left[A_1^c\right]P\left[A_2\right]. \tag{1.5.4}$$

The proofs of the other two conclusions are similar and are left as an exercise.

The binomial distribution arises when N independent observations are made on events A and A^c belonging to some fixed probability space $(\Omega, \mathcal{A}, P)$. At each observation, we are interested in the occurrence of the event A or its complement A^c. For example, at the birth of a child we would be interested in the complementary events "the child is a boy" or "the child is a girl", and the birth of successive children in a family may be considered as independent observations on these events . For the sake of brevity set $P[A] = p$ and $P[A^c] = q$, where $0 \leq p \leq 1$ and $q = 1 - p$.

Making N independent observations on the events A and A^c induces another probability space $(\Omega_1, \mathcal{A}_1, P_1)$, which may be constructed according to the following scheme. At each observation, we shall record a 1 if the event A occurs, and a 0 if its complement A^c occurs. The set Ω_1 may be defined as the set of all N dimensional sequences of the form $101\ldots1$ containing some combination of zeros and ones, *i.e.*, every $\omega \in \Omega_1$ will be such a sequence. For example, if $N = 4$, then 1011 would be the sequence for which the event A occurred at the first, third, and fourth observations and the event A^c occurred at the second observation.

Since each place in every N dimensional sequence may be filled by either a zero or a one, it is clear that the set Ω_1 contains 2^N points. The algebra $\mathcal{A}_1$ may be taken as the class $\mathbb{C}$ of all sets in Ω_1. It is of interest to note that the class $\mathbb{C}$ will contain a very large number

$$2^{2^N} \tag{1.5.5}$$

of sets even for moderate values of N. For example, for $N = 10$

$$2^{2^{10}} = 1.797\,693\,134\,862\,32 \times 10^{308}, \tag{1.5.6}$$

which indeed is a very large number. As a matter of fact, the algebra $\mathcal{A}_1$ may contain many more sets than are actually of interest.

A probability P_1 may be defined on $\mathcal{A}_1$ by proceeding as follows. If a sequence $\omega \in \Omega_1$ contains x ones and $N-x$ zeroes, then under the assumption of independence, the event (ω) would be assigned the probability

$$P_1\left[(\omega)\right] = p^x q^{N-x}. \tag{1.5.7}$$

For every event E belonging to $\mathcal{A}_1$, put

$$P_1\left[E\right] = \sum_{\omega \in E} P_1\left[(\omega)\right]. \tag{1.5.8}$$

It is easy to check that the resulting triple $(\Omega_1, \mathcal{A}_1, P_1)$ is a finite probability space.

The event of greatest interest in this discussion is the event that A is observed x times in N independent observations. This event may be easily characterized in terms of N functions $\xi_k\left(\omega\right)$ for $k = 1, 2, \ldots, N$, called random variables, which are defined on Ω_1 and take values in the set consisting of the numbers 0 and 1. More precisely, these functions are defined on Ω_1 as follows. For each index $k = 1, 2, \ldots, N$ set $\xi_k\left(\omega\right) = 0$ or $\xi_k\left(\omega\right) = 1$ if the k *-th* position in the sequence ω contains a 0 or 1, respectively. Such random variables are also known as Bernoulli indicators. Given this definition, the number of times the event A occurs in N independent observations is given by the random variable

$$Z_N\left(\omega\right) = \sum_{k=1}^{N} \xi_k\left(\omega\right). \tag{1.5.9}$$

The possible values of the random variable $Z_N\left(\omega\right)$ are $0, 1, 2, \ldots, N$, and the event that A occurs x times in N independent observations or trials will be denoted by

$$\left[\omega \mid Z_N\left(\omega\right) = x\right], \tag{1.5.10}$$

where this symbol is read as the set of $\omega \in \Omega_1$ such that $Z_N\left(\omega\right) = x$. For the sake of brevity, we shall frequently drop the ω and write $\left[Z_N = x\right]$ for this set. Note that as $x = 0, 1, 2, \ldots, N$ the events in (1.5.10) form a partition of the set Ω_1.

The next task in the development of ideas underlying the binomial distribution is to derive an expression for the probability of the event $\left[Z_N = x\right]$. By the assumption of independence, every $\omega \in \left[Z_N = x\right]$ is assigned the probability $p^x q^{N-x}$. Therefore,

$$P_1\left[\left[Z_N = x\right]\right] = \sum_{\omega \in \left[Z_N = x\right]} p^x q^{N-x}. \tag{1.5.11}$$

The desired formula for this probability may be derived as soon as the number of points in $[Z_N = x]$ is determined. But, $\omega \in [Z_N = x]$ if, and only if, x ones are contained among the N zeroes and ones in the sequence ω. Hence, the number of points in $[Z_N = x]$ is the number of ways one can choose x places in the sequence from N places, which is the number $\binom{N}{x}$. Hence,

$$f(x) = P_1\left[[Z_N = x]\right] = \binom{N}{x} p^x q^{N-x} \tag{1.5.12}$$

for $x = 0, 1, 2, \ldots, N$.

The function $f(x)$ in (1.5.12) is known as the probability density function for the binomial distribution. Observe that if $p > 0$, then $f(x) > 0$ for all $x = 0, 1, 2, \ldots, N$. Furthermore, it can be seen that

$$\sum_{x=0}^{N} f(x) = \sum_{x=0}^{N} \binom{N}{x} p^x q^{N-x} = (p+q)^N = 1\ . \tag{1.5.13}$$

In subsequent chapters the binomial distribution will be encountered so that it is desirable to have some convenient notation for stating that some random variable X has a binomial distribution with parameters N and p. From now on, if a random variable X is assumed to have a binomial distribution, then we shall write

$$X \sim B(N, p)\,, \tag{1.5.14}$$

which reads X has a binomial distribution with index or sample size N and probability p. Usually, N will be a positive integer and p is a number such that $0 < p < 1$.

Having defined the binomial distribution, we are now ready for a few applications of this distribution in genetics.

Example 1.5.1: Distribution of Boys and Girls in Families of Size *N*

A concrete example of the kinds of questions that may be answered in terms of the binomial distribution is the following. In families with 5 children, what is the probability of having 3 boys and 2 girls?

If it is assumed that the birth of children in a family are independent observations on the complementary events, “the child is a boy” or “the child is a girl”, then the binomial distribution applies. If it is further assumed that these complementary events have equal probabilities $p = q = \frac{1}{2}$, then the required probability is

$$\frac{5!}{3!2!}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^2 = \frac{5!}{3!2!}\left(\frac{1}{2}\right)^5 = \frac{5}{16}. \tag{1.5.15}$$

Actually, data suggests that the probability $p = 105/205$ so that $q = 100/205$, but for the sake of simplicity this information was ignored in the example.

Example 1.5.2: Offspring From a Mating of Type $Aa \otimes Aa$

Suppose allele A is dominant to allele a. In four offspring from a mating of type $Aa \otimes Aa$, what is the probability of observing three of the dominant phenotype $A-$ and one offspring of the recessive genotype aa?

If again it is assumed that the binomial distribution applies, then the probability of observing the dominant phenotype is $p = 3/4$ so that $q = 1/4$. Therefore, the required probability is

$$\frac{4!}{3!1!}\left(\frac{3}{4}\right)^3\frac{1}{4} = \frac{27}{64}\ . \tag{1.5.16}$$

This is the probability of getting a perfect $3:1$ ratio of the dominant to the recessive phenotypes in the mating $Aa \otimes Aa$. It is of interest to note that the probability of not getting this perfect ratio is 37/64. This simple example is typical of a more general situation. That is, when a large number of offspring are observed from a mating of the type under consideration, the probability of not observing a perfect $3:1$ ratio is usually higher than that of observing a perfect ratio.

Example 1.5.3: Observing at Least One Mutation with High Probability

Mutations are rare events with small probabilities. For example, suppose a bacterium of type A mutates to one of type B with probability p, where p is in the range $10^{-10} \le p \le 10^{-6}$. Assume that in a population of N bacteria, mutations occur independently among the members of this population so that whether at least one mutation occurs in the population may be described by a binomial distribution. Let E denote the event that at least one mutation is observed in the population of N individual cells.

Then, formally this event is the union

$$E = \sum_{x=1}^{N} [Z_N = x] , \tag{1.5.17}$$

and this event occurs with probability

$$P_1 [E] = \sum_{x=1}^{N} \binom{N}{x} p^x q^{N-x} . \tag{1.5.18}$$

The problem to be considered is that, given a value of p, find a value of N such that

$$P_1 [E] = 0.95. \tag{1.5.19}$$

Let E^c denote the complement of E. Then,

$$P_1 [E^c] = (1-p)^N = 0.05 \tag{1.5.20}$$

so that the problem of finding N such that $P_1 [E] = 0.95$ reduces to solving for N equation (1.5.20) . From this equation, it can be seen that

$$N = \left[\frac{\ln (0.05)}{\ln (1-p)} \right] , \tag{1.5.21}$$

where $[\cdot]$ stands for the greatest integer in a number. A call to the computation engine for the case $p = 10^{-6}$ results in the value

$$\frac{\ln (0.05)}{\ln (1 - 10^{-6})} = 2,995,730.\,775\,687\,6 \tag{1.5.22}$$

so that the estimate of N is $2,995,730$. Similarly, for $p = 10^{-10}$

$$\frac{\ln (0.05)}{\ln (1 - 10^{-10})} = 2.\,995\,732\,273\,404\,2 \times 10^{10} . \tag{1.5.23}$$

Therefore, in this case N is of order 3×10^{10}. In conclusion, for $p = 10^{-10}$, an investigator would need to grow a rather large population of bacteria to find at least one mutation with probability 0.95.

1.6 The Multinomial Distribution

Having introduced the binomial distribution, we are now ready for its multidimensional generalization, the multinomial distribution. The multinomial distribution is even more widely applicable in genetics than the binomial distribution, since it characterizes situations in which more than two disjoint events may be observed at each independent observation. As a simple example of this situation in genetics consider the ABO blood system in man. This system is evidently controlled by three alleles denoted by A, B and O at a single locus. Actually, there seems to be more than three alleles at this locus but for the sake of simplicity only three alleles will be considered. Individuals of genotypes AA and AO have type A blood, those with genotypes BB and BO have type B blood, those of genotype OO have type O blood, and those of genotype AB have type AB blood. In terms of Mendelian genetics, both alleles A and B are dominant to allele O, but A and B are codominant.

From a mating of the type $AB \otimes AB$, an offspring may have any one of three blood types; namely A, B, or AB, and we may be interested in the probability of the event each blood type is represented in families of size three from such matings. Or, in general, in families of size $N \geq 1$ from a mating of type $AB \otimes AB$, what is the probability of observing x_1, x_2 and x_3 of genotypes AA, AB and BB, respectively, where $x_1 + x_2 + x_3 = N$? Questions of this type may be answered within the framework of the multinomial distribution.

In general, let $(\Omega, \mathcal{A}, P)$ be any probability space and for $r \geq 2$ let $A_1 \ldots., A_r$ be events in $\mathcal{A}$, which form a partition of the set Ω. That is,

$$\sum_{k=1}^{r} A_k = \Omega. \tag{1.6.1}$$

Let $P\left[A_k\right] = p_k$, where $p_k \geq 0$ and

$$\sum_{k=1}^{r} p_k = 1. \tag{1.6.2}$$

Just as was the case of the multinomial distribution, making N independent observations on the r disjoint events $A_1, \ldots, A_r$ induces another probability space$(\Omega_1, \mathcal{A}_1, P_1)$ which may be constructed according to the following scheme.

Let $\varepsilon_1, \varepsilon_2, \ldots . \varepsilon_r$ denote a set of r-dimensional vectors such that for $k = 1, 2, \ldots, r$, the vector ε_k has a one in the *k-th* position and zeros elsewhere. If $r = 3$, for example, then this set of vectors would be $\varepsilon_1 = (1, 0, 0)$, $\varepsilon_2 = (0, 1, 0)$, and $\varepsilon_3 = (0, 0, 1)$. In what follows, these vectors will be used as indicators to record which one of the r events occurs at each independent trial. If, for example, event A_k occurs, then the occurrence of this event will be denoted by ε_k. A sequence of N independent observations will be denoted by a sequence of the form $\varepsilon_{i_1}\varepsilon_{i_2} \ldots \varepsilon_{i_N}$, where for i_ν for $\nu = 1, 2, \ldots, N$, $i_v \in (1, 2, \ldots, r)$. If, for example, $N = 4$, then the symbol $\varepsilon_1\varepsilon_2\varepsilon_3\varepsilon_4$ represents the case events A_1, A_2, A_3, A_4 occurred on independent trials 1,2,3, and 4, respectively.

For a fixed N, the set Ω_1 may be taken as the set of all sequences ω of $\boldsymbol{\varepsilon}'s$ containing N such symbols. It is clear, therefore, that the set Ω_1 contains r^N elements. The algebra $\mathcal{A}_1$ may be taken as the set $\mathbb{C}$, the class of all subsets of the set Ω_1. It is interesting to note that the class $\mathbb{C}$ contains

$$2^{r^N} \tag{1.6.3}$$

sets, which in general will be a very large number. A probability function P_1 may be defined on $\mathcal{A}_1$, using the following procedure. If a point ω contains the symbol $\boldsymbol{\varepsilon}_k$ at x_k places in the sequence ω, where $x_1 + x_2 + \ldots + x_r = N$, then, under the assumption that the N trials are independent, the set (ω) would be assigned the probability

$$P_1\left[(\omega)\right] = p_1^{x_1} p_2^{x_2} \ldots p_r^{x_r}. \tag{1.6.4}$$

Then, for any event $E \in \mathcal{A}_1$, the probability $P_1\left[E\right]$ would be assigned the value

$$P_1\left[E\right] = \sum_{\omega \in E} P_1\left[(\omega)\right] \tag{1.6.5}$$

It is of interest to note that the sum on the right contains only finitely many terms. Given these definitions, it is easy to verify that the triple $(\Omega_1, \mathcal{A}_1, P_1)$ is a finite probability space.

We are particularly interested in the event that A_k occurs $x_k \geq 0$ times in N independent trials, where $x_1 + x_2 + \cdots + x_r = N$. This event is easily characterized in terms of N generalized Bernoulli indicator random variables $\boldsymbol{X}_1(\omega), \boldsymbol{X}_2(\omega), \ldots, \boldsymbol{X}_N(\omega)$ defined for $\omega \in \Omega_1$ and taking values in the set $(\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \ldots . \boldsymbol{\varepsilon}_r)$ of r-dimensional indicator vectors. If $\omega \in \Omega_1$ is a sequence containing the vector $\boldsymbol{\varepsilon}_v$ for $\nu = 1, 2, \ldots, r$ in the *k-th* position, then set

$$\boldsymbol{X}_k(\omega) = \boldsymbol{\varepsilon}_\nu. \tag{1.6.6}$$

It will be noted that by using this definition a unique value is assigned to each of the vector valued indicator functions $\boldsymbol{X}_1(\omega), \ldots, \boldsymbol{X}_N(\omega)$ for every $\omega \in \Omega_1$.

For $\nu = 1, 2, \ldots, r$, let $Z_{\nu N}(\omega)$ denote the number of times event A_ν occurs in N independent trials and define a r-dimensional vector valued random variable $\boldsymbol{Z}_N(\omega)$ by

$$\boldsymbol{Z}_N(\omega) = \left(\boldsymbol{Z}_{1N}(\omega), \ldots, \boldsymbol{Z}_{rN}(\omega)\right) \tag{1.6.7}$$

for all $\omega \in \Omega_1$. It can be seen that

$$\boldsymbol{Z}_N(\omega) = \boldsymbol{X}_1(\omega) + \boldsymbol{X}_2(\omega) + \cdots + \boldsymbol{X}_N(\omega), \tag{1.6.8}$$

where the sum on the right is the sum of N vectors each of $r \geq 2$ dimensions. The summation of vectors is defined as element by element summation. For example, if $\boldsymbol{a} = (a_1, a_2, \ldots, a_r)$ and $\boldsymbol{b} = (b_1, b_2, \ldots, b_r)$, then $\boldsymbol{a} + \boldsymbol{b} = (a_1 + b_1, a_2 + b_2, \ldots, a_r + b_r)$. Moreover, $\boldsymbol{a} = \boldsymbol{b}$ if, and only if, $a_\nu = b_\nu$ for $\nu = 1, 2, \ldots, r$.

The event that A_ν occurs x_v times in N independent trials is the set of $\omega \in \Omega_1$ such that $\boldsymbol{Z}_N(\omega) = \boldsymbol{x}$, where $\boldsymbol{x} = (x_1, x_2, \ldots, x_r)$ and $x_1 + x_2 + \cdots + x_r = N$. In symbols, this event is denoted by

$$[\omega \mid \boldsymbol{Z}_N(\omega) = \boldsymbol{x}] \tag{1.6.9}$$

but to lighten the notation we shall write $[\boldsymbol{Z}_N = \boldsymbol{x}]$. By definition, the probability of this event is

$$P_1[\boldsymbol{Z}_N = \boldsymbol{x}] = \sum_{\omega \in [\boldsymbol{Z}_N = \boldsymbol{x}]} p_1^{x_1} p_2^{x_2} \ldots p_r^{x_r}. \tag{1.6.10}$$

Therefore, to derive a formula for this probability it will be necessary to count the number of points in the set $[\boldsymbol{Z}_N = \boldsymbol{x}]$. However, this number is the number of ways that a set containing N elements can be partitioned into sets $A_1, A_2, \ldots, A_r$, where set A_ν contains x_ν elements and $x_1 + x_2 + \cdots + x_r = N$. As was seen in section (1.4) this number is

$$\frac{N!}{x_1! \times x_2! \times \ldots \times x_r!}. \tag{1.6.11}$$

Therefore, the desired formula is

$$f(\boldsymbol{x}) = P_1[\boldsymbol{Z}_N = \boldsymbol{x}] = \frac{N!}{x_1! x_2! \ldots x_r!} p_1^{x_1} p_2^{x_2} \ldots p_r^{x_r}. \tag{1.6.12}$$

This formula is known as the probability density function for the multinomial distribution with sample size or index N and probability vector $(p_1, p_2, \ldots, p_r)$.

The number of terms in this formula as the $\boldsymbol{x}$ varies over all solutions in the set

$$\mathfrak{S} = [\boldsymbol{x} \mid x_1 + x_2 + \cdots + x_r = N] \tag{1.6.13}$$

is very large, and in section (1.4) we saw that the number of points in the set $\mathfrak{S}$ was

$$\binom{N + r - 1}{N} \tag{1.6.14}$$

for $r \geq 2$. According to the multinomial theorem, it follows that

$$\sum_{\boldsymbol{x}} f(\boldsymbol{x}) = (p_1 + p_2 + \cdots + p_r)^N = 1, \tag{1.6.15}$$

where the sum runs over all $\boldsymbol{x} \in \mathfrak{S}$.

We are now in a position to give an answer to the question posed at the beginning of this section as well as answers to many other questions.

Example 1.6.1: A, B, AB and O Blood Types - Matings of Type $\boldsymbol{AB \otimes AB}$

In families of size 3 from matings of type $AB \otimes AB$, what is the probability of observing offspring with each of the three blood types A, B and AB?.

Offspring with these blood types are produced in matings of this type with probabilities 1/4.1/4 and 1/2, respectively. Therefore, by applying formula (1.6.12) , it follows that

$$\frac{3!}{1!1!1!}\left(\frac{1}{4}\right)\left(\frac{1}{4}\right)\left(\frac{1}{2}\right) = \frac{3}{16}. \tag{1.6.16}$$

Observe that the probability that at least one blood type is not represented is 13/16.

It is also of interest to consider the sex of individuals in mating of the type under consideration as in the following example.

Example 1.6.2: Sex and Blood Types in Mating of Type $\boldsymbol{ABXX \otimes ABXY}$

In mating of the type under consideration a female symbolized on the left will produce the gametes AX and BX with equal probabilities of 1/2. A male on the right in such matings will produce gametes AX, BX, AY and BY with equal probabilities of 1/4. Now imagine constructing a 2 × 4 table with eight cells, representing the eight kinds of genotypes that may be present in the offspring of such matings. As blood type in this case is controlled by genes at an autosomal locus and is, therefore, independent of the inheritance of sex, each of these eight cells will be assigned a probability of 1/8. Given this assignment of probabilities, the set of probabilities

$$\begin{aligned} P\,[\text{Female Type } A] &= P\,[\text{Male Type } A] = \frac{1}{8} \\ P\,[\text{Female Type } B] &= P\,[\text{Male Type } B] = \frac{1}{8} \\ P\,[\text{Female Type } AB] &= P\,[\text{Male Type } AB] = \frac{2}{8} \end{aligned} \tag{1.6.17}$$

may be easily derived. In such families of size six, what is the probability of observing a boy and a girl of each of the three blood types? This is another case in which it is assumed that the multinomial distribution applies. Thus, the answer to this questions is

$$6!\left(\frac{1}{8}\right)^2\left(\frac{1}{8}\right)^2\left(\frac{2}{8}\right)^2=\frac{45}{4096}. \tag{1.6.18}$$

The production of offspring from matings of the type under consideration may be viewed as placing balls into six cells according to the probabilities stated above. The six cells in this case correspond to the six possible genotypes that may be produced by a mating under consideration. Under this interpretation the number in (1.6.18) is the probability that all cells are occupied. The probability of the complementary event "that at least one cell is not occupied" is 4051/4096, which is much larger than the probability that all cells are occupied.

This last example is, again, typical of a general situation. In populations in which several alleles are present at a large number of loci, most matings are capable of producing a large number of genotypes. Yet in any family only a few of the possible genotypes will be represented, for let r be the number of genotypes possible from a mating, let N be the family size, and consider the event all genotypes are represented in the offspring from the mating. If $N < r$, the event has probability zero, and if $N > r$, then the event will have positive probability but will in general have small probability whenever N and r are of the same order of magnitude.

If we extend our considerations to the population as a whole in which several alleles may be present at each of a large number of loci, then the possible number of genotypes generated by the population will in general be enormous. At any time the composition of a population of size N may be visualized as N balls placed in some random fashion into a large number r of cells, which represent the possible genotypes. As the population evolves in time, different sets of cells will be occupied, but at any fixed time relatively few cells will be occupied, particularly when $N < r$. In a manner of speaking, one of the main tasks of mathematical genetics is to describe and analyze mathematically the biological factors influencing the sets of cells occupied during the course of evolution of a population. For a further discussion of occupancy problems see Feller (1968) Chapter II.

When constructing a finite probability space, the set Ω of basic events may be constructed in a variety of ways. For example, rather than using the approach described in this section, one could proceed by defining probabilities on the set $\mathfrak{S}$ in (1.6.13) directly. There is, for example, a distribution in physics known as the Maxwell-Blotzmann distribution, where it is assumed that

$$p_\nu = \frac{1}{r} \tag{1.6.19}$$

for all $\nu = 1, 2, \ldots, r$. In this case, for any $\boldsymbol{x} \in \mathfrak{S}$,

$$f(\boldsymbol{x}) = P[\boldsymbol{Z}_B = \boldsymbol{x}] = \frac{N!}{x_1!x_2!\ldots x_r!}\frac{1}{r^N}. \tag{1.6.20}$$

There is also a distribution in physics, known as the Bose-Enstein distribution, in which it is assumed that $\boldsymbol{x} \in \mathfrak{S}$

$$f(\boldsymbol{x}) = P[\boldsymbol{Z}_B = \boldsymbol{x}] = \frac{1}{\binom{N+r-1}{N}}. \tag{1.6.21}$$

As it turns out, however, neither of these distributions fit in all experiments in which the movements of physical particles such as nuclei of atoms, electrons and protons were viewed as placing N balls at random into r cells. From such experience, one reaches the conclusion that it is very difficult to assign probabilities to events in a Ω which may subsequently be confirmed in experiments. An interested reader may again consult Feller (1968). In a subsequent section of this chapter, we shall return to the problem of verifying some proposed probability model empirically.

1.7 Conditional Probabilities and a Bayesian Theorem

The idea of conditional probability plays an important role in mathematical genetics and other applications of probability. As a point of reference, let $(\Omega, \mathcal{A}, P)$ denote some finite probability space under consideration, and let A and B denote two events in $\mathcal{A}$. Then, if $P[A] \neq 0$, the conditional probability of the event B, given the event A, is defined as

$$P[B \mid A] = \frac{P[AB]}{P[A]}. \tag{1.7.1}$$

If $AB = \varphi$, then $P[B \mid A] = 0$.

To illustrate this idea, suppose $\Omega = (\omega \mid \omega = 1, 2, 3, \ldots, 10), A = (2, 3, 4, 5)$ and $B = (4, 5, 6)$. Then, $AB = (4, 5)$ so that $P[AB] = 2/10$ and $P[A] = 4/10$. Thus, in this example,

$$P[B \mid A] = \frac{\frac{2}{10}}{\frac{4}{10}} = \frac{1}{2}\,. \tag{1.7.2}$$

The conditional probability $P[A \mid B]$ is defined similarly for those cases for which $P[B] \neq 0$.

Given these definitions, it follows that $P\left[AB\right]$ may be expressed in two ways, namely,

$$P\left[AB\right]=P\left[A\right]P\left[B\mid A\right]=P\left[B\right]P\left[A\mid B\right], \tag{1.7.3}$$

whenever the probabilities on the right are well defined. More generally, if C is another event in $\mathcal{A}$, then $P\left[ABC\right]$ may be expressed as

$$P\left[ABC\right]=P\left[A\right]P\left[B\mid A\right]P\left[C\mid AB\right], \tag{1.7.4}$$

whenever the probabilities on the right are well defined. It is of interest to note that by defining additional conditional probabilities, the probability $P\left[ABC\right]$ may be expressed in $3!=6$ ways. And, in general, if the intersection of n events $A_1,A_2,\ldots,A$ in $\mathcal{A}$ is under consideration, then the probability

$$P\left[\bigcap_{k=1}^{n}A_k\right] \tag{1.7.5}$$

may be expressed in $n!$ ways, which can be a very large number even for moderate values of n. Whether all or some of these conditional probabilities are well defined would depend on how probabilities were assigned to the finitely many points in $\Omega=(\omega_i\mid i=1,2,\ldots,N)$, where N is an integer such that $N\geq 1$.

A formula involving conditional probabilities, known as Bayes' theorem, arises frequently in mathematical genetics and other applications of probability and statistics. For $r\geq 2$, let B_k, for $k=1,2,\ldots,r$, be sets in $\mathcal{A}$ which are a partition of the set Ω so that for any $A\in\mathcal{A}$

$$A=A\cap\Omega=\sum_{k=1}^{r}AB_k. \tag{1.7.6}$$

Therefore,

$$P\left[A\right]=\sum_{k=1}^{r}P\left[AB_k\right]=\sum_{k=1}^{r}P\left[B_k\right]P\left[A\mid B_k\right]. \tag{1.7.7}$$

For any $k=1,2,\ldots,r$ and $P\left[A\right]>0$, consider the conditional probability

$$P\left[B_k\mid A\right]=\frac{P\left[AB_k\right]}{P\left[A\right]}=\frac{P\left[B_k\right]P\left[A\mid B_k\right]}{\sum_{k=1}^{r}P\left[B_k\right]P\left[A\mid B_k\right]}. \tag{1.7.8}$$

In much of the literature on probability and statistics, this formula is known as Bayes formula or theorem.

A problem that frequently emerges when attempting to formulate a mathematical model which accommodates the idea of natural selection in evolutionary genetics is that of expressing the action of selection in terms of quantitative probabilities. The following is an illustrative example of a situation in evolutionary genetics in which Bayes' theorem may be applied.

Example 1.7.1: Natural Selection with Respect to One Autosomal Locus with Two Alleles

Let C and c denote two alleles at some autosomal locus, and let the symbols B_1, B_2 and B_3, respectively, denote the three genotypes CC, Cc and cc that may be present in a population in any generation $n \geq 1$. Then, $P\left[B_k\right]$ for $k = 1, 2, 3$ would be interpreted as the probability of finding genotype B_k in the population in generation n. In mathematical genetics these probabilities are also called genotypic frequencies. Let A denote the event that selection acts on the three genotypes in the population and let $P\left[A \mid B_k\right]$ be the conditional probability that genotype B_k remains in the population in generation $n+1$ after the action of natural selection and let $P\left[B_k \mid A\right]$ be the probability or frequency of genotype B_k in the population in generation $n+1$ after the action of natural selection. Then, by Bayes theorem this conditional probability is given by

$$P\left[B_k \mid A\right] = \frac{P\left[B_k\right] P\left[A \mid B_k\right]}{\sum_{k=1}^{3} P\left[B_k\right] P\left[A \mid B_k\right]}, \tag{1.7.9}$$

for $k = 1, 2, 3$.

An alternative way of expressing this formula in a language and notation that is more widely used in evolutionary genetics is to suppose that a population of diploid organisms reproduce in discrete time generations and let $Q_k\left(n\right)$ for $k = 1, 2, 3$ denote, respectively, the frequencies of the genotypes CC, Cc and cc in generation $n \geq 1$. Let v_k denote the probability that genotype k is still present in the population after selection occurs. Then, by an application of Bayes theorem formula, the frequency of the genotypes that would produce the offspring in generation $n+1$ would be

$$Q_k\left(n+1\right) = \frac{Q_k\left(n\right) v_k}{\sum_{k=1}^{3} Q_k\left(n\right) v_k}, \tag{1.7.10}$$

for $k = 1, 2, 3$. In some subsequent chapters, the concept of natural selection will be discussed and formulated extensively in terms in the language of probability.

1.8 Expectations and Generating Functions for Binomial and Multinomial Distributions

The concept of expectations of random variables is one of tools that is used frequently in mathematical genetics and various fields of probability and statistics in developing the properties of model. In this section examples of this concept will be given in terms of the binomial and multinomial distributions. As in section 1.5, let the random variable X have a binomial distribution with index $N \geq 1$ and probability p, where $0 < p < 1$. Then, the probability density function of X is

$$P_1 [X = x] = f(x) = \binom{N}{x} p^x (1-p)^{N-x} \tag{1.8.1}$$

for $x = 0, 1, 2, \ldots, N$. The expectation of X is defined by the equation

$$E[X] = \sum_{x=0}^{N} x f(x) = \sum_{x=0}^{N} x \binom{N}{x} p^x (1-p)^{N-x} . \tag{1.8.2}$$

In what follows the substitution $q = 1 - p$ will be used. By manipulating the formula on the right algebraically, it is possible to reduce the sum to a simple formula, but such manipulations can be rather tedious.

For the case of a binomial distribution, the idea of a generating function for a discrete distribution reduces the problem of deriving formulas for expectations by using operations involving calculus. For s such that $0 \leq s \leq 1$, the generating function of the random variable X is defined by the expectation

$$\begin{aligned} g(s) = E\left[s^X\right] &= \sum_{x=0}^{N} s^x \binom{N}{x} p^x (1-p)^{N-x} \\ &= \sum_{x=0}^{N} \binom{N}{x} (ps)^x (1-p)^{N-x} = (ps+q)^N \end{aligned} \tag{1.8.3}$$

by an application of the binomial theorem. Now suppose we differentiate this formula with respect to s to obtain

$$\frac{dg(s)}{ds} = \sum_{x=0}^{N} x s^{x-1} \binom{N}{x} p^x (1-p)^{N-x} = Np (ps+q)^{N-1} . \tag{1.8.4}$$

By putting $s = 1$ in this equation, it can be seen that

$$E[X] = \sum_{x=0}^{N} x \binom{N}{x} p^x (1-p)^{N-x} = Np = \frac{dg(1)}{ds} . \tag{1.8.5}$$

To lighten the notation let $\mu = Np$
The variance of the random variable X is defined by the expectation

$$\begin{aligned} var\left[X\right] &= E\left[\left(X-\mu\right)^2\right] = \sum_{x=0}^{N}\left(x-\mu\right)^2\binom{N}{x}p^x\left(1-p\right)^{N-x} \\ &= \sum_{x=0}^{N} x^2\binom{N}{x}p^x\left(1-p\right)^{N-x} - \mu^2 \\ &= E\left[X^2\right] - \mu^2. \end{aligned} \tag{1.8.6}$$

In the derivation of this formula, we have tacitly used the definition

$$E\left[X^2\right] = \sum_{x=0}^{N} x^2\binom{N}{x}p^x\left(1-p\right)^{N-x}. \tag{1.8.7}$$

To derive a formula for the variance of X, it will be necessary to find a formula for this expectation.
To this end, suppose we differentiate equation (1.8.4) with respect to s. Then,

$$\frac{d^2g\left(s\right)}{ds^2} = \sum_{x=0}^{N} x\left(x-1\right)s^{x-1}\binom{N}{x}p^x\left(1-p\right)^{N-x} = N\left(N-1\right)p^2\left(ps+q\right)^{N-2}. \tag{1.8.8}$$

By setting $s = 1$ in this equation, it can be seen that

$$E\left[X\left(X-1\right)\right] = \sum_{x=0}^{N} x\left(x-1\right)\binom{N}{x}p^x\left(1-p\right)^{N-x} = N\left(N-1\right)p^2. \tag{1.8.9}$$

However,

$$E\left[X^2\right] = E\left[X\left(X-1\right)\right] + E\left[X\right] = N\left(N-1\right)p^2 + Np. \tag{1.8.10}$$

Therefore,

$$\begin{aligned} var[X] &= N\left(N-1\right)p^2 + Np - N^2p^2 \\ &= Np - Np^2 = Np\left(1-p\right) = Npq. \end{aligned} \tag{1.8.11}$$

Similar expectations may be defined for the multinomial distribution. To illustrate the concepts consider the case $r = 3$. In this case, the vector random variable $\boldsymbol{X} = \left(X_1, X_2, X_3\right)$ has the probability density function

$$f\left(x_1, x_2, x_3\right) = \frac{N!}{x_1!x_2!x_3!}p_1^{x_1}p_2^{x_2}p_3^{x_3}, \tag{1.8.12}$$

where $x_1 + x_2 + x_3 = N \geq 1$, $p_{1+}p_{2+}p_3 = 1$ and $p_k \geq 0$ for $k = 1, 2, 3$. Just as in section $\left(1.5\right)$, let $\mathfrak{S}$ denote the set of all solutions of the equations

$x_1 + x_2 + x_3 = N$, where $x_k \geq 0$ for $k = 1, 2, 3$. Then, the generating function of the multinomial distribution for the case $r = 3$ is defined by the expectation

$$\begin{aligned} g(s_1, s_2, s_3) &= E\left[s_1^{X_1} s_2^{X_2} s_3^{X_3}\right] = \sum_{\mathbf{x} \in \mathfrak{S}} \frac{N!}{x_1! x_2! x_3!} (s_1 p_1)^{x_1} (s_2 p_2)^{x_2} (s_3 p_3)^{x_3} \\ &= (p_1 s_1 + p_2 s_2 + p_3 s_3)^N . \end{aligned} \tag{1.8.13}$$

The expectation of the random variable X_1 is defined by the equation

$$E[X_1] = \sum_{\mathbf{x} \in \mathfrak{S}} x_1 \frac{N!}{x_1! x_2! x_3!} p_1^{x_1} p_2^{x_2} p_3^{x_3}. \tag{1.8.14}$$

To evaluate this sum symbolically, consider the partial derivative

$$\frac{\partial g(s_1, s_2, s_3)}{\partial s_1} = N p_1 (p_1 s_1 + p_2 s_2 + p_3 s_3)^{N-1}. \tag{1.8.15}$$

By evaluating this equation at $s_k = 1$ for $k = 1, 2, 3$ and using an argument similar to that for the binomial distribution, one can show that the desired expectation is

$$E[X_1] = N p_1. \tag{1.8.16}$$

In general, it can be shown that $E[X_k] = N p_k$ for $k = 1, 2, 3$. By using a derivation similar to that for the binomial distribution, it can also be shown that the variances of the random variables are given by the formula

$$var[X_k] = N p_k (1 - p_k) \tag{1.8.17}$$

for $k = 1, 2, 3$.

For the case of two or more random variables, there is also the concept of covariances, which are measures of covariation among a set of random variables. For the case of the random variables X_1 and X_2, their covariance is defined by the expectation

$$\begin{aligned} cov[X_1, X_2] &= E[(X_1 - E[X_1])(X_2 - E[X_2])] \\ &= E[X_1 X_2] - E[X_1] E[X_2]. \end{aligned} \tag{1.8.18}$$

Just as was the case for the binomial distribution, a formula for the expectation $E[X_1 X_2]$ may be derived by differentiating equation (1.8.15) with respect to s_2. Thus,

$$\frac{\partial^2 g(s_1, s_2, s_3)}{\partial s_1 \partial s_2} = N(N-1) p_1 p_2 (p_1 s_1 + p_2 s_2 + p_3 s_3)^{N-2} . \tag{1.8.19}$$

By evaluating this equation at $s_k = 1$ for $k = 1, 2, 3$, it can be seen that

$$E[X_1 X_2] = N(N-1) p_1 p_2. \tag{1.8.20}$$

Therefore, the covariance of X_1 and X_2 is

$$cov[X_1, X_2] = N(N-1) p_1 p_2 - N^2 p_1 p_2 = -N p_1 p_2. \tag{1.8.21}$$

In general, for $j \neq k$, $cov[X_j, X_k] = -N p_j p_k$.

1.9 Marginal and Conditional Distributions of the Multinomial Distribution

In subsequent chapters, the concepts of marginal and conditional distributions associated with some multivariate distribution will be needed. In this section, these concepts will be illustrated using the multinomial distribution. For the case $r = 3$, the density function of the multinomial distribution may be expressed in the form

$$f(x_1, x_2) = \binom{N}{x_1}\binom{N - x_1}{x_2} p_1^{x_1} p_2^{x_2} (1 - p - p_2)^{N - x_1 - x_2}, \tag{1.9.1}$$

for $x_1 = 0, 1, 2, \ldots, N$ and for each x_1, $x_2 = 0, 1, \ldots, N - x_1$. In deriving this formula from (1.8.12), we have used the conditions $x_1 + x_2 + x_3 = N \geq 1$, $p_1 + p_2 + p_3 = 1$ and $p_k \geq 0$ for $k = 1, 2, 3$. Observe that this formula is the joint density of the random variables X_1 and X_2.

By definition, the marginal density function of the random variable X_1 is the sum

$$f_1(x_1) = \sum_{x_2=0}^{N-x_1} f(x_1, x_3) \tag{1.9.2}$$

for each fixed $x_1 = 0, 1, 2, \ldots, N$. In terms of the joint density function in (1.9.1), this function may be expressed as

$$\begin{aligned} f_1(x_1) &= \binom{N}{x_1} p_1^{x_1} \sum_{x_2=0}^{N-x_1} \binom{N - x_1}{x_2} p_2^{x_2} (1 - p_1 - p_2)^{N - x_1 - x_2} \\ &= \binom{N}{x_1} p_1^{x_1} (1 - p_1)^{N - x_1} \end{aligned} \tag{1.9.3}$$

for $x_1 = 0, 1, 2, \ldots, N$, by an application of the binomial theorem. It can also be shown, by using a similar derivation, that the marginal density of the random variable X_2 is

$$f_2(x_2) = \binom{N}{x_2} p_2^{x_2} (1 - p_2)^{N - x_2} \tag{1.9.4}$$

for $x_2 = 0, 1, 2, \ldots, N$. From (1.9.3) and (1.9.4), it may be concluded that both the random variables X_1 and X_2 have binomial distributions.

With regard to the concept of a conditional distribution, it should be pointed out that when dealing with random variables the intersection of two or more sets is expressed in a different way than that used in previous sections. Let $(\Omega_1, \mathcal{A}_1, P_1)$ denote the probability space defined in section

1.6 underlying the multinomial distribution. Then, in the notation of that section

$$[X_1 = x_1, X_2 = x_2] = [X_1 = x_1] \cap [X_2 = x_2], \tag{1.9.5}$$

which reads $[X_1 = x_1, X_2 = x_2]$ is the intersection of sets $[X_1 = x_1]$ and $[X_2 = x_2]$. Given this notation, the conditional probability of the event $[X_2 = x_2]$, given the event $[X_1 = x_1]$, is defined as

$$f(x_2 \mid x_1) = \frac{P_1 [X_1 = x_1, X_2 = x_2]}{P_1 [X_1 = x_1]} = \frac{f(x_1, x_2)}{f_1(x_1)}. \tag{1.9.6}$$

The function $f(x_2 \mid x_1)$ is known as the conditional density of X_2, given $X_1 = x_1$. From (1.9.1) and (1.9.4), it can be seen that

$$\begin{aligned} f(x_2 \mid x_1) &= \frac{\binom{N}{x_1}\binom{N-x_1}{x_2} p_1^{x_1} p_2^{x_2} (1-p_1-p_2)^{N-x_1-x_2}}{\binom{N}{x_1} p_1^{x_1} (1-p_1)^{N-x_1}} \\ &= \frac{\binom{N-x_1}{x_2} p_2^{x_2} (1-p_1-p_2)^{N-x_1-x_2}}{(1-p_1)^{N-x_1}} \\ &= \binom{N-x_1}{x_2} \left(\frac{p_2}{1-p_1}\right)^{x_2} \left(1 - \frac{p_2}{1-p_1}\right)^{N-x_1-x_2} \end{aligned} \tag{1.9.7}$$

for $x_2 = 0, 1, 2, \ldots, N - x_1$. In this last step of this derivation, the observation that $(1-p_1)^{N-x_1} = (1-p_1)^{x_2} (1-p_1)^{N-x_1-x_2}$ was used. Therefore, the conditional distribution of the random variable X_2, given $X_1 = x_1$, is a binomial distribution with index $N - x_1$ and probability $p_2/(1-p_1)$.

When working with the multinomial distribution, the easiest way of finding marginal distribution is by inspection of the generating function after setting some of the variables in the function equal to 1. To illustrate this idea, consider the equation

$$\begin{aligned} g(s_1, s_2) &= \sum_{x_1=0}^{N} \sum_{x_2=0}^{N-x_1} \binom{N}{x_1} \binom{N-x_1}{x_2} (p_1 s_1)^{x_1} (p_2 s_2)^{x_2} (1-p_1-p_2)^{N-x_1-x_2} \\ &= (p_1 s_1 + p_2 s_2 + 1 - p_1 - p_2)^N \end{aligned} \tag{1.9.8}$$

for the generating function of the probability density function in (1.9.1). If we set $s_2 = 1$ in this equation, then the equation

$$\begin{aligned} g_1(s_1) &= g(s_1, 1) \\ &= \sum_{x_1=0}^{N} \sum_{x_2=0}^{N-x_1} \binom{N}{x_1} \binom{N-x_1}{x_2} (p_1 s_1)^{x_1} (p_2)^{x_2} (1-p_1-p_2)^{N-x_1-x_2} \\ &= (p_1 s_1 + p_2 + 1 - p_1 - p_2)^N = (p_1 s_1 + 1 - p_1)^N \end{aligned} \tag{1.9.9}$$

arises. From this equation, it can be seen that $(p_1 s_1 + 1 - p_1)^N$ is the generating function of the binomial density given in $(1.9.3)$. Given this result, we could deduce directly that the marginal density of the random variable X_1 is that for the binomial distribution without resorting to doing the symbolic summation in $(1.9.3)$.

In general, for $r \geq 1$ and $N \geq 1$, if the vector random variable $\boldsymbol{X} = (X_1, X_2, \ldots, X_r)$ has a multinomial distribution with probabilities p_k for $k = 1, 2, \ldots, r$, then its generating function is

$$g(s_1, s_2, \ldots, s_r) = (p_1 s_1 + p_2 s_2 + \cdots + p_r s_r)^N . \tag{1.9.10}$$

Thus, if we let $s_k = 1$ for $k = 2, 3, \ldots, r$, then it can be seen that

$$g_1(s_1) = (p_1 s_1 + p_2 + \cdots + p_r)^N = (p_1 s_1 + 1 - p_1)^N . \tag{1.9.11}$$

From this result, it can be concluded that the density function of the random variable X_1 is that for the binomial distribution in $(1.9.3)$. In general, by using this technique, one can conclude that the density function of the random variables X_k for $k = 2, \ldots, r$ is that for the binomial distribution with index N and probability p_k.

For the case of two random variables X_1 and X_2, if $s_k = 1$ for $k = 3, \ldots, r$ in (1.9.10), then it can be seen that the generating function of the pair (X_1, X_2) is

$$g_{12}(s_1, s_2) = (p_1 s_1 + p_2 s_2 + 1 - p_1 - p_2)^N \tag{1.9.12}$$

so that it may be concluded that the pair (X_1, X_2) has the density function displayed in $(1.9.1)$. More generally, let $(X_{i_1}, X_{i_2}, \ldots, X_{i_k})$ denote a subvector of the vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_r)$ of dimension k such that $1 \leq k < N$, then by using the technique just described, it can be concluded that the subvector $(X_{i_1}, X_{i_2}, \ldots, X_{i_k})$ has a multinomial distribution with index parameter N and probabilities p_{i_ν} for $\nu = 1, 2, \ldots, k$. If $s_k = \exp(t_k)$, where t_k is any real number, for $k = 1, 2, \ldots, r$, then the function that would result if these symbols were substituted into a generating function is called the moment generating function. Moreover, if $s_k = \exp(iu_k)$, where $i^2 = -1$ and u_k is real, for $k = 1, 2, \ldots, r$, then the complex valued function that would result if these symbols were substituted into a generating function is called the characteristic function. For these functions, marginal distributions could be derived by letting $t_{i_\nu} = 0$ and $u_{i_\nu} = 0$ for $\nu = 1, 2, \ldots, k$, at some subset $(i_{v_1}, \ldots, i_{\nu_k})$ in $(1, 2, \ldots, r)$.

Given these results on deriving marginal distributions of the multinomial distribution, it is also possible to extend the result in (1.9.7) for the conditional distribution of the random variable X_2, given that $X_1 = x_1$. To simplify the notation, if a random variable X has a binomial distribution with index parameter N and probability p, then we shall write

$$X \sim B\left(N, p\right) \tag{1.9.13}$$

as in section 1.5. The symbol $X_2 \mid X_1 = x_1$ will also be used to denote the conditional distribution of X_2, given that $X_1 = x_1$. In this notation, the result stated in (1.9.7) may be written as

$$X_2 \mid X_1 = x_1 \sim B\left(N - x_1, \frac{p_2}{1 - p_1}\right). \tag{1.9.14}$$

Similarly, the symbol $X_3 \mid X_1 = x_1, X_2 = x_2$ will denote the conditional distribution of X_3, given that $X_1 = x_1$ and $X_2 = x_2$. In this notation, it can be shown that

$$X_3 \mid X_1 = x_1, X_2 = x_2 \sim B\left(N - x_1 - x_2, \frac{p_3}{1 - p_1 - p_2}\right). \tag{1.9.15}$$

The density function of this conditional distribution will be denoted by $f\left(x_3 \mid x_1, x_2\right)$. By continuing to derive conditional distributions in this way, it can be shown that in general the density of the multinomial distribution may be factored into a product of conditional densities of the form

$$\begin{aligned} & f(x_1, x_2, \ldots, x_r) \qquad (1.9.16) \\ &= f_1\left(x_1\right) f\left(x_2 \mid x_1\right) f\left(x_3 \mid x_1, x_2\right) \ldots f\left(x_r \mid x_1, \ldots, x_{r-1}\right). \end{aligned}$$

As will be shown in a subsequent section, this factorization is particularly useful in designing algorithms to compute Monte Carlo realizations of a vector random variable $\boldsymbol{X}$ with a multinomial distribution.

1.10 A Law of Large Numbers and the Frequency Interpretation of Probability

There is another approach to finding expectations and variances of binomial random variables that will be useful in subsequent chapters of this book. Moreover, this approach is very useful in defining the concept of the law of large numbers within the frequency interpretation of probability. As was seen in section 1.5 in equation (1.5.9), for example, a random variable with a binomial distribution with index N and probability p may be represented in the form

$$X\left(\omega\right) = Z_N\left(\omega\right) = \sum_{k=1}^{N} \xi_k\left(\omega\right), \tag{1.10.1}$$

where $\xi_k(\omega)$ are independent Bernoulli indicator random variables such that $\xi_k(\omega) = 1$ with probability p and $\xi_k(\omega) = 0$ with probability $q = 1-p$ for $k = 1, 2, \ldots, N$. Observe that, by definition, $E[\xi_k(\omega)] = 1 \times p + 0 \times q = p$ for all $k = 1, 2, \ldots, N$.

It is also useful to observe that the probability density function of any Bernoulli indicator random variable $\xi_k(\omega)$ has the form

$$P[\xi_k(\omega) = x_k] = f_k(x_k) = p^{x_k}(1-p)^{1-x_k} \tag{1.10.2}$$

for $x_k = 0, 1$ and $k = 1, 2, \ldots, N$. Moreover, for the case $N = 2$, the statement $\xi_1(\omega)$ and $\xi_2(\omega)$ are independent random variables means that their joint density function is

$$f_{12}(x_1, x_2) = P[[\xi_1(\omega) = x_1, \xi_2(\omega) = x_2]] = p^{x_1}(1-p)^{1-x_1} p^{x_2}(1-p)^{1-x_2}, \tag{1.10.3}$$

for $x_1 = 0, 1$ and $x_2 = 0, 1$. In this case, the expectation of the random variable $Z_2(\omega)$ is, by definition,

$$\begin{aligned} E[Z_2(\omega)] &= E[\xi_1(\omega) + \xi_2(\omega)] \\ &= \sum_{x_1=0}^{1} \sum_{x_2=0}^{1} (x_1 + x_2) f_{12}(x_1, x_2) \\ &= \sum_{x_1=0}^{1} x_1 f_1(x_1) + \sum_{x_2=0}^{1} x_2 f_2(x_2) \\ &= E[\xi_1(\omega)] + E[\xi_2(\omega)] = 2p. \end{aligned} \tag{1.10.4}$$

For $N \geq 2$ it can be shown, by extending the argument just given, that

$$\begin{aligned} E[X(\omega)] = E[Z_N(\omega)] &= E\left[\sum_{k=1}^{N} \xi_k(\omega)\right] \\ &= \sum_{k=1}^{N} E[\xi_k(\omega)] = Np, \end{aligned} \tag{1.10.5}$$

as was shown by another method in section 1.8.

By proceeding as in (1.10.5) for the case $N = 2$, it can be shown that

$$var[Z_2(\omega)] = var[\xi_1(\omega)] + var[\xi_2(\omega)] + 2cov[\xi_1(\omega), \xi_2(\omega)]. \tag{1.10.6}$$

But, as was indicated in section 1.8,

$$cov[\xi_1(\omega), \xi_2(\omega)] = E[\xi_1(\omega)\xi_2(\omega)] - E[\xi_1(\omega)]E[\xi_2(\omega)]. \tag{1.10.7}$$

Because $E[\xi_1(\omega)] = E[\xi_2(\omega)] = p$, it follows that $E[\xi_1(\omega)]E[\xi_2(\omega)] = p^2$. Furthermore, from the definition of expectation, it can be seen, by the independence assumption, that

$$E[\xi_1(\omega)\xi_2(\omega)] = 1 \times P[\xi_1(\omega) = 1, \xi_2(\omega) = 1] = p^2. \tag{1.10.8}$$

Hence, $cov\left[\xi_1(\omega), \xi_2(\omega)\right] = 0$. It can also be shown that $var\left[\xi_1\right] = var\left[\xi\right] = p(1-p)$. Therefore,

$$var\left[Z_2(\omega)\right] = 2p(1-p), \tag{1.10.9}$$

and in general, by using the ideas just outlined, it can be shown that

$$var\left[Z_N(\omega)\right] = Np(1-p). \tag{1.10.10}$$

for $N \geq 2$ as was shown in section 1.8.

The random variable $Z_N(\omega)$ is the number of times some event A under consideration is observed in $N \geq 1$ independent Bernoulli trials. Therefore, the random variable

$$f_N(\omega) = \frac{1}{N} Z_N(\omega) = \frac{1}{N} \sum_{k=1}^{N} \xi_k(\omega) \tag{1.10.11}$$

is the frequency or fraction of times the event A occurs in N independent Bernoulli trials. The probability that the event A occurs in any trial is p and it is assumed that p is unknown. According to the frequency interpretation of the concept of probability, the unknown number probability p is the fraction of times the event A would be observed in a large number N of independent Bernoulli trials. It thus seems plausible to interpret $f_N(\omega)$ as an estimate of p.

Because $f_N(\omega)$ is a random variable, it will vary among realizations ω of an experiment. It is, therefore, of interest to develop some properties of the random variable $f_N(\omega)$. From the definition of this random variable, it can be seen that

$$E\left[f_N(\omega)\right] = \frac{1}{N} E\left[Z_N(\omega)\right] = \frac{1}{N} Np = p. \tag{1.10.12}$$

This result seems very desirable, because it shows us that if we average the random variable $f_N(\omega)$ over all conceivable experiments, then this average is p, the unknown probability. It is also of interest to get a measure of the variability of the fraction $f_N(\omega)$ around its expectation p. By definition, a measure of this variation is the variance

$$var\left[f_N(\omega)\right] = E\left[\left(f_N(\omega) - p\right)^2\right] = \frac{1}{N^2} Np(1-p) = \frac{p(1-p)}{N}. \tag{1.10.13}$$

This is also an interesting result, because for any fixed value $p \in (0,1)$ of the unknown probability p, the variance $var\left[f_N(\omega)\right]$ is small when N is large.

Equation (1.9.13) is informative, but it is also of interest to frame a statement of a law of large numbers in terms of a probability statement.

More precisely, let $\epsilon > 0$ be a small number close to zero. Then, to express a law of large numbers in terms of a probability statement, consider the probability

$$P_1 \left[\mid f_N(\omega) - p \mid > \epsilon \right] \tag{1.10.14}$$

for N large. Observe, that this probability statement is with respect to the probability space $(\Omega_1, \mathcal{A}_1, P_1)$ constructed in section 1.5, where the binomial distribution was defined within that framework. In terms of this distribution, this probability may be expressed as

$$P_1 \left[\mid f_N(\omega) - p \mid > \epsilon \right] = \sum_{z \in D} \binom{N}{z} p^z (1-p)^{N-z}, \tag{1.10.15}$$

where D is the set $D = [z \mid z/N - p \mid > \epsilon]$. At this point in the discussion, the problem that arises is that of finding a way to prove that the probability on the right in (1.10.15) is small for every $\epsilon > 0$ when N is large.

As will be shown in what follows, a solution of this problem may be found by applying what is known as Chebychev's inequality. To derive this inequality, consider the variance

$$var\left[f_N(\omega)\right] = \sum_{z=0}^{N} (z/N - p)^2 \binom{N}{z} p^z (1-p)^{N-z}. \tag{1.10.16}$$

This sum may be expressed in terms of two sums as follows:

$$var\left[f_N(\omega)\right] = \sum_{z \in D^c} (z/N - p)^2 f(z) + \sum_{z \in D} (z/N - p)^2 f(z), \tag{1.10.17}$$

where

$$f(z) = \binom{N}{z} p^z (1-p)^{N-z}$$

to lighten the notation. From this equation, it can be seen that

$$\begin{aligned} var\left[f_N(\omega)\right] &\geq \sum_{z \in D} (z/N - p)^2 f(z) \\ &\geq \epsilon^2 \sum_{z \in D} f(z) \\ &= \epsilon^2 P_1 \left[\mid f_N(\omega) - p \mid > \epsilon \right]. \end{aligned} \tag{1.10.18}$$

Therefore,

$$0 \leq P_1 \left[\mid f_N(\omega) - p \mid > \epsilon \right] \leq \frac{1}{\epsilon^2} var\left[f_N(\omega)\right] = \frac{p(1-p)}{\epsilon^2 N}, \tag{1.10.19}$$

which is known as Chebychev's inequality for the binomial distribution. Actually, this inequality applies to any distribution with a finite variance.

From this inequality, it can be seen that for every $\epsilon > 0$,

$$\lim_{N \uparrow\uparrow \infty} P_1 \left[\mid f_N (\omega) - p \mid > \epsilon \right] = 0 \tag{1.10.20}$$

for every p such that $0 < p < 1$. This statement is a precise way to express the frequency interpretation of probability in the sense that the probability of the complementary event

$$P_1 \left[\mid f_N (\omega) - p \mid \leq \epsilon \right] \to 1 \tag{1.10.21}$$

is close to one as $N \uparrow \infty$ for every $\epsilon > 0$, which gives precise meaning to the statement that the frequency $f_N (\omega)$ is close to p with probability close to one, when the sample size N is large.

In Mendelian genetics, the law of large numbers is often applied tacitly without any mention of its mathematical basis. For example, suppose an autosomal gene A is dominant to its allele a. Then, among the offspring from a mating of type $Aa \otimes Aa$, it is expected that the phenotype $A-$ would be observed with probability $3/4$. In practice what this statement usually means is the following. Let N denote a large number of offspring from a mating of the type under consideration and let $n(A-)$ denote the number of offspring that were classified as phenotype $A-$. Then, the frequency of this phenotype among the offspring is $n(A-)/N \simeq 3/4$, where $\simeq$ means approximately. When an experimenter makes such statement, a law of large numbers and the frequency interpretation of probability is being tacitly applied. In advanced books on probability theory, the concept of laws of large numbers are treated in depth with distributions other than the binomial, but such treatments will not be considered here.

1.11 On Computing Monte Carlo Realizations of a Random Variable with a Binomial Distribution

As will be illustrated in subsequent chapters, in population genetics complex stochastic models of evolutionary processes of high dimensionality arise that are not amenable to analysis, using classical methods of mathematics. When such models are entertained, investigators often resort to Monte Carlo simulation methods in which realizations of a stochastic process are computed and statistically analyzed to deduce interesting substantive properties of the model under investigation that are useful in interpreting observed data. Monte Carlo simulation methods center around programming a computer to compute independent realizations of random variables that have a uniform distribution on the set of integers $(x \mid x = 1, 2, \ldots, M)$, where M is a large positive integer.

By definition, a random variable X taking values in this set of integers is said to have a uniform distribution if its probability density function has the form

$$f(x) = \frac{1}{M} \tag{1.11.1}$$

for $x = 1, 2, \ldots, M$. Of course, one may entertain uniform distributions on any finite set of objects that are not integers. For instance, in the examples from Mendelian genetics that were discussed in foregoing sections of this chapter, it was often assumed that the *a prori* distribution on some set of gametes or genotypes produced by a mating under consideration was uniform, i.e., all points in the set were assigned equal probabilities.

As a first step in the development of ideas to simulate realizations of a random variable with a binomial distribution, let $(\Omega, \mathcal{A}, P)$ denote some probability space under consideration and suppose event $A \in \mathcal{A}$ has probability $P[A] = p$. Let $\xi(\omega)$ denote a Bernoulli indicator of the event A. Because p belongs to the interval $(0, 1)$, to simulate that event A occurs, a random variable taking values in the interval $(0, 1)$ will be needed. This need will be met if we define a random variable U_M by

$$U_M = \frac{X}{M}, \tag{1.11.2}$$

which takes values in the set $(u \mid u = 1/M, 2/M, \ldots, 1)$. Let $U_M = u$ denote a computed realization of the random variable U_M. Then, if $u \leq p$, set $\xi(\omega) = 1$, indicating that the event A occurred, and if $u > p$, set $\xi(\omega) = 0$, indicating the event A did not occur.

As was seen in equation (1.5.9), a random variable $Z_N(\omega)$ with a binomial distribution may be represented in the form

$$Z_N(\omega) = \sum_{k=1}^{N} \xi_k(\omega), \tag{1.11.3}$$

which is the sum of N independent Bernoulli indicators. Therefore, a solution to the problem of simulating realizations of a random variable with a binomial distribution reduces to that of simulating N independent realizations of Bernoulli indicators $\xi_k(\omega)$ for $k = 1, 2, \ldots, N$ by repeating the procedure outlined above. In this simulation, the word, independent, is a key word so that when programming a computer to simulate realizations of the random variable U_M, one should test statistically whether the simulated realizations of the random variable do indeed satisfy the condition of independence. The problem of designing random number generators will be discussed in more detail in a subsequent chapter.

When N is large, computing a realization of the random variable $Z_N(\omega)$ in (1.11.3) will entail many calls to a random number generator, which can lead to long execution times to complete a Monte Carlo simulation experiment. When N is large and p is small, a case that will frequently arise when an investigator wishes to simulate genetic mutations, the binomial distribution may be approximated by the Poisson distribution. To this end, let $\lambda = Np$ and write the binomial density function in the form

$$\begin{aligned} f_N(x) &= \frac{N(N-1)\ldots(N-k+1)}{x!}\left(\frac{\lambda}{N}\right)^x\left(1-\frac{\lambda}{N}\right)^{N-x} \\ &= \left(1-\frac{1}{N}\right)\ldots\left(1-\frac{k+1}{N}\right)\frac{\lambda^x}{x!}\left(1-\frac{\lambda}{N}\right)^{N-x}. \end{aligned} \tag{1.11.4}$$

From this expression, it can be seen that

$$\lim_{N\uparrow\infty} f_N(x) = g(x) = e^{-\lambda}\frac{\lambda^x}{x!} \tag{1.11.5}$$

for all $x = 0, 1, 2, \ldots$. This formula is the density function of a random variable with a Poisson distribution. Given this formula, the problem of computing a realization of a random variable with a binomial distribution, when N is large and p is small, reduces the problem of computing a realization of a random variable with a density function in (1.11.5) from a realization of the random variable U_M. Procedures for carrying out such calculations will also be discussed in a subsequent chapter.

Cases also arise when N is large and p is not small when designing procedures for simulating realizations of a binomial random variable. In such cases, the binomial distribution may be approximated by a normal distribution. A random variable Z is said to have a standard normal distribution if its probability density function has the form

$$\phi(z) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right), \tag{1.11.6}$$

where $-\infty < z < \infty$. The distribution function of the random variable Z is defined as the integral

$$P(Z \le z) = \Phi(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}\exp\left(-\frac{s^2}{2}\right)ds, \tag{1.11.7}$$

where z is a real number in $-\infty < z < \infty$.

Let V_N denote a random variable defined by the equation

$$V_N = \frac{Z_N - Np}{\sqrt{Np(1-p)}}, \tag{1.11.8}$$

where Z_N is a binomial random variable with expectation Np and variance $Np(1-p)$. It can be seen that $E[V_N] = 0$ and $var[V_N] = 1$. It is proved in books on mathematical probability and statistics that

$$\lim_{N\uparrow\infty} F_N(z) = \lim_{N\uparrow\infty} P[V_N \leq z] = \Phi(z) \tag{1.11.9}$$

for all v such that $-\infty < v < \infty$. This result is known as the central limit theorem for the binomial distribution.

From this result, one may conclude that

$$\frac{Z_N - Np}{\sqrt{Np(1-p)}} \approx Z, \tag{1.11.10}$$

where the symbol $\approx$ means approximately in distribution. An equivalent way of writing this relationship is

$$Z_N \approx Np + (\sqrt{Np(1-p)})Z. \tag{1.11.11}$$

Thus, when N is large and p is not small, the problem of computing a realization of a binomial random variable Z_N reduces to the problem of computing a realization of a standard normal random variable Z from a realization of the random variable U_M. Problems of this type will also be discussed in a subsequent chapter.

When applying relation (1.11.11), two caveats must be observed. The random variable Z_N in (1.11.11) may, in theory, take on any value z such that $-\infty < z < \infty$ but Z_N takes on the integer values $0, 1, 2, \ldots, N$. Applications of formula (1.11.11) will work well when Np is sufficiently large so that negative values of Z_N will occur with small probability. To ensure that computed realizations of the random variable Z_N are not negative and take on integer values, it will often be sufficient to use the formula

$$Z_N = \left[\mid Np + (\sqrt{Np(1-p)})Z \mid\right], \tag{1.11.12}$$

where the function $\mid \cdot \mid$ stands for the absolute value and the function $[\cdot]$ denotes the greatest integer in x, where x is any non-negative real number. To ensure that realized values of Z_N are $z = 0, 1, 2, \ldots, N$, a computer may be programmed to accept Z_N if it is in the set $(z \mid 0, 1, 2, \ldots, N)$ but if $Z_N > N$, set $Z_N = N$.

Finally, from equation (1.9.16), it can be seen that if a vector random variable $\boldsymbol{Z}_N = (Z_{N1}, \ldots, Z_{Nr})$ has a multinomial distribution with index

$N \geq 1$ and probability vector $\boldsymbol{p} = (p_1, p_2, \ldots, p_r)$, then it suffices to compute realizations of binomial random variable recursively with conditional distributions indicated on the righthand side of that equation.

1.12 The Beta-Binomial Distribution

In the foregoing sections of this chapter, the probability p in the binomial distribution has been viewed as a constant, which may be either known or unknown. Whenever one considers a model of observable phenomena, rather than stopping with some particular formulation, it seems advisable to consider at least one alternative formulation so that the properties of the formulations may be compared. In this section, therefore, this probability will be viewed as a realization of a continuous type random variable $\Theta = \theta$, taking values in the interval $(0, 1)$.

Let $g_1(\theta)$ denote the probability density function of the random variable Θ. The idea that the probability θ is a realization of a random variable may be interpreted in various ways in genetic applications. For example, suppose in families of size $N \geq 1$, it is of interest to collect data on the number of boys and girls observed in a sample of such families. In some of these families, the data may suggest that the birth of a boy is more probable than that of a girl, but in other families girls may be more prevalent. Such observations suggest that it would be of interest to entertain a model in which the probability θ that a baby is a boy is a realization of a random variable Θ that may vary among families in the sample.

As a first step in formulating such a model, let Z_N denote a random variable taking values in the set $(z \mid z = 0, 1, 2, \ldots, N)$, and suppose the conditional density of Z_N, given $\Theta = \theta$, is the binomial density

$$g(z \mid \theta) = \binom{N}{z} \theta^z (1-\theta)^{N-z} \tag{1.12.1}$$

for $z = 0, 1, 2, \ldots, N$. Then, by using an extension of the concept of conditional probabilities discussed in section (1.7), it can be seen that the joint density function of the random variable Θ and Z_N is defined as

$$g(z, \theta) = g_1(\theta)\, g(z \mid \theta) \tag{1.12.2}$$

for $z = 0, 1, 2, \ldots, N$ and $\theta \in (0, 1)$. The marginal or unconditional density of the random variable Z_N is, by definition,

$$f(z) = \int_0^1 g(z, \theta)\, d\theta = \int_0^1 g_1(\theta)\, g(z \mid \theta)\, d\theta. \tag{1.12.3}$$

To complete the formulation of the model, the next step is that of making a judicious choice of $g_1(\theta)$, the probability density function of the random variable Θ. A useful choice for this function is one that has properties such that the integral in (1.12.3) may be expressed in terms of well known special functions of mathematics.

Such a well known function is the gamma function, which is defined by the integral

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx. \tag{1.12.4}$$

It can be shown that this integral converges for all $\alpha > 0$ so that the function is well defined for all $\alpha > 0$. It can also be shown by an integration by parts, that this function satisfies the recursive equation

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha). \tag{1.12.5}$$

In particular, if $\alpha = n$, a positive integer, then

$$\Gamma(n+1) = n\Gamma(n) = n(n-1)\Gamma(n-1) = n!, \tag{1.12.6}$$

because

$$\Gamma(1) = \int_0^\infty e^{-x} dx = 1. \tag{1.12.7}$$

One approach to considering the gamma function is to note that, although it is a continuous function of $\alpha > 0$, it coincides with the factorial function when α is a positive integer. From recursive equation (1.12.5), it also follows that for every integer $n \geq 1$

$$\Gamma(\alpha+n) = (\alpha+n-1)(\alpha+n-2)\ldots(\alpha+1)\alpha\Gamma(\alpha), \tag{1.12.8}$$

which will be useful in what follows.

The beta function $B(\alpha,\beta)$ is defined by the integral

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx. \tag{1.12.9}$$

It can be shown that this integral converges for all $\alpha > 0$ and $\beta > 0$. This integral is sometimes called Euler's first integral. As it turns out, it can be shown that

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \tag{1.12.10}$$

whenever $\alpha > 0$ and $\beta > 0$. For further details on the gamma and beta functions, the book Artin (1964) may be consulted or the names of the gamma and beta functions may be entered into a search engine for the world wide web, where a considerable amount of information may also be found.

Given equation (1.12.10), a random variable Θ taking values in the interval $(0,1)$ is said to have a beta distribution if its probability density function is

$$g_1(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, \tag{1.12.11}$$

where $\theta \in (0,1)$. From this definition, it can be seen that the expectation of the random variable Θ is

$$\begin{aligned} E[\Theta] &= \int_0^1 \theta g_1(\theta)\,d\theta = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 \theta^{\alpha}(1-\theta)^{\beta-1}\,d\theta \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+1+\beta)} = \frac{\alpha}{\alpha+\beta}. \end{aligned} \tag{1.12.12}$$

Similarly, although the derivation is more involved, it can be shown that the variance of Θ is

$$\begin{aligned} var[\Theta] &= E\left[\Theta^2\right] - (E[\Theta])^2 \\ &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \end{aligned} \tag{1.12.13}$$

As we shall see, selecting the beta distribution for the probability density function of the random variable Θ is a judicious choice for the density $g_1(\theta)$, because the integral in (1.12.3) may be expressed in terms of gamma functions. Thus, the unconditional density of the random variable Z_N is

$$\begin{aligned} f(z) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 \binom{N}{z}\theta^{\alpha+z-1}(1-\theta)^{\beta+N-z-1}\,d\theta \\ &= \binom{N}{z}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+z)\Gamma(\beta+N-z)}{\Gamma(\alpha+\beta+N)} \end{aligned} \tag{1.12.14}$$

for $z = 0, 1, 2, \ldots, N$. From (1.12.8), it can be seen that for every integer $n \geq 1$

$$\frac{\Gamma(\alpha+n)}{\Gamma(\alpha)} = \alpha(\alpha+1)(\alpha+2)\ldots(\alpha+n-1). \tag{1.12.15}$$

To simplify the notation, let the symbol $\alpha^{(n)}$ denote the righthand side of this equation. Formally, this function could be defined by the ratio

$$\alpha^{(n)} = \frac{\Gamma(\alpha+n)}{\Gamma(\alpha)}, \tag{1.12.16}$$

and from this definition, it can be seen that $\alpha^{(0)} = 1$ so that $\alpha^{(n)}$ is well defined for all $n = 0, 1, 2, \ldots$. Given these definitions, from an inspection of (1.12.14), it can be seen that the density for the beta-binomial density may be represented in the succinct form

$$f(z) = \binom{N}{z} \frac{\alpha^{(z)}\beta^{(N-z)}}{(\alpha+\beta)^{(N)}} \tag{1.12.17}$$

for $z = 0, 1, 2, \ldots, N$. When an investigator wishes to compute the numerical version of the density, this symbolic form is especially useful in writing software to compute the values $f(z)$, given numerical values of the parameters N, α and β, for $z = 0, 1, 2, \ldots, N$.

Symbolic forms of the unconditional expectation and variance of the random variable Z_N are not only of intrinsic interest but can also be useful when estimating values of the unknown parameters α and β from data. In principle, these symbolic forms can be derived from the formula in (1.12.17) by evaluating the expressions

$$E[Z_N] = \sum_{z=0}^{N} z f(z) \tag{1.12.18}$$

and

$$var[Z_N] = E\left[(Z_N - E[Z_N])^2\right] = \sum_{z=0}^{N} (z - E[Z_N])^2 f(z) \tag{1.12.19}$$

symbolically, but this approach leads to difficult exercises in symbol manipulations. A simpler approach is to derive a formula for the expectation directly, using properties of conditional expectations.

As the conditional distribution of the random variable Z_N is binomial with index N and random probability Θ, it follows that the conditional expectation of Z_N, given Θ, is

$$E[Z_N \mid \Theta] = N\Theta. \tag{1.12.20}$$

By definition, the unconditional expectation Z_N is the expectation of this conditional expectation with respect to the distribution of the random variable Θ. In symbols, this expectation takes the form

$$E[Z_N] = E[E[Z_N \mid \Theta]] = NE[\Theta] = N\frac{\alpha}{\alpha+\beta}, \tag{1.12.21}$$

by (1.12.12).

As a first step in deriving a symbolic form for the unconditional variance of the random variable Z_N, consider the equation

$$Z_N - N\frac{\alpha}{\alpha+\beta} = (Z_N - N\Theta) + N\left(\Theta - \frac{\alpha}{\alpha+\beta}\right). \tag{1.12.22}$$

By squaring both sides of this equation, it can be seen that

$$\begin{aligned}\left(Z_N - N\frac{\alpha}{\alpha+\beta}\right)^2 &= \left(Z_N - N\Theta\right)^2 \\ &\quad + N^2\left(\Theta - \frac{\alpha}{\alpha+\beta}\right)^2 \\ &\quad + 2\left(N\left(\Theta - \frac{\alpha}{\alpha+\beta}\right)\right)\left(Z_N - N\Theta\right). \qquad (1.12.23)\end{aligned}$$

By taking the conditional expectation of this equation, given Θ, it can be seen that

$$\begin{aligned}& E\left[\left(Z_N - N\frac{\alpha}{\alpha+\beta}\right)^2 \mid \Theta\right] \\ &= E\left[\left(Z_N - N\Theta\right)^2 \mid \Theta\right] \\ &\quad + N^2 E\left[\left(\Theta - \frac{\alpha}{\alpha+\beta}\right)^2 \mid \Theta\right] \\ &\quad + 2E\left[\left(N\left(\Theta - \frac{\alpha}{\alpha+\beta}\right)\right)\left(Z_N - N\Theta\right) \mid \Theta\right]. \qquad (1.12.24)\end{aligned}$$

However,

$$E\left[\left(Z_N - N\frac{\alpha}{\alpha+\beta}\right)^2 \mid \Theta\right] = N\Theta\left(1-\Theta\right), \qquad (1.12.25)$$

because the conditional expectation on the left is with respect to a binomial distribution with index N and probability Θ. Moreover,

$$N^2 E\left[\left(\Theta - \frac{\alpha}{\alpha+\beta}\right)^2 \mid \Theta\right] = N^2\left(\Theta - \frac{\alpha}{\alpha+\beta}\right)^2, \qquad (1.12.26)$$

because the expression on the right is a function only of Θ. In general, a little reflection will show that $E\left[h\left(X\right) \mid X\right] = h\left(X\right)$, because a function of X acts as if it were a constant when taking a conditional expectation, given X. By using this observation again, it can be seen that

$$\begin{aligned}& 2E\left[\left(N\left(\Theta - \frac{\alpha}{\alpha+\beta}\right)\right)\left(Z_N - N\Theta\right) \mid \Theta\right] \\ &= 2\left(N\left(\Theta - \frac{\alpha}{\alpha+\beta}\right)\right)E\left[\left(Z_N - N\Theta\right) \mid \Theta\right] = 0, \qquad (1.12.27)\end{aligned}$$

because $E[(Z_N - N\Theta) \mid \Theta] = N\Theta - N\Theta = 0$.

Therefore, the unconditional variance of the random variable Z_N has the form

$$\begin{aligned} var\left[Z_N\right] &= E\left[\left(Z_N - N\frac{\alpha}{\alpha+\beta}\right)^2\right] \\ &= NE\left[\Theta\left(1-\Theta\right)\right] + N^2E\left[\left(\Theta - \frac{\alpha}{\alpha+\beta}\right)^2\right]. \end{aligned} \tag{1.12.28}$$

However,

$$\begin{aligned} E\left[\Theta\left(1-\Theta\right)\right] &= \frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}\int_0^1 \theta^{\alpha}\left(1-\theta\right)^{\beta} d\theta \\ &= \frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}\frac{\Gamma\left(\alpha+1\right)\Gamma\left(\beta+1\right)}{\Gamma\left(\alpha+\beta+2\right)} \\ &= \frac{\alpha\beta\left(\alpha+\beta\right)}{\left(\alpha+\beta\right)^2\left(\alpha+\beta+1\right)}. \end{aligned} \tag{1.12.29}$$

Moreover,

$$E\left[\left(\Theta - \frac{\alpha}{\alpha+\beta}\right)^2\right] = var\left[\Theta\right]. \tag{1.12.30}$$

Thus, by using (1.12.13) and by substituting (1.12.29) into (1.12.28), it can be shown that

$$var\left[Z_N\right] = \frac{N\alpha\beta\left(\alpha+\beta+N\right)}{\left(\alpha+\beta\right)^2\left(\alpha+\beta+1\right)}. \tag{1.12.31}$$

In deriving this formula, a well known general result, involving conditional expectations and variances, has been utilized. By using the techniques similar to those outlined above, it can be shown that, in general, when two random variables X and Y are under consideration,

$$var\left[Y\right] = E\left[var[Y \mid X]\right] + var\left[E\left[Y \mid X\right]\right]. \tag{1.12.32}$$

In subsequent chapters, this formula will be used when various models of stochastic population dynamics are under consideration.

As mentioned in the beginning of this section, it would be of interest to consider both the binomial and beta-binomial distributions as competing models for the distribution of the number of boys in families with some fixed number of children. In this connection Derman, Gleser, and Olkin (1973) have presented data on the number of boys for $M = 53,680$ German families with 8 children, see page 307 of their book. From such data, one can estimate the empirical distribution of the number of boys in families of

8 children without reference to any preconceived model and test whether the binomial or beta-binomial yields the best fit to the data as will be illustrated in the example below.

Example 1.12.1: Empirical Distribution of the Number of Boys in German Families with 8 Children

With regard to the German family data, let $M\left(x\right)$ denote the number of families with $z = 0, 1, 2, \ldots, 8$ boys and let M denote the total number of families observed. Then, an empirical estimate of the density function of the number of boys in such families is

$$\widehat{f}\left(z\right) = \frac{M\left(z\right)}{M} \tag{1.12.33}$$

for $z = 0, 1, 2, \ldots, 8$. Given this function, the mean number of boys in such families would be estimated by

$$\overline{Z} = \sum_{z=0}^{8} z\widehat{f}\left(z\right), \tag{1.12.34}$$

and an estimate of the variance of the number of boys would be given by

$$\widehat{\sigma}^2 = \sum_{z=0}^{8} (z - \overline{Z})^2 \widehat{f}\left(z\right). \tag{1.12.35}$$

An empirical estimate of the probability p that a child is a boy in families with 8 children, would be given by

$$\widehat{p} = \frac{1}{8}\overline{Z}. \tag{1.12.36}$$

To verify that $\widehat{p}$ is the frequency of boys in the data, observe that $M\overline{Z}$ is the total number of boys in the sample. Altogether, there are $8M$ children in the sample. Hence, the frequency of boys is $M\overline{Z}/8M = \overline{Z}/8$. If the binomial distribution does indeed fit the data, then $\overline{Z} \simeq 8\widehat{p}$ and $\widehat{\sigma}^2 \simeq 8\widehat{p}\,(1 - \widehat{p})$, where $\simeq$ means approximately. Actually, from (1.12.36), it follows that $\overline{Z} = 8\widehat{p}$.

Rather than relying solely on comparing $\overline{Z}$ and $\widehat{\sigma}^2$ with theoretical expectation and variance, it is advisable to test the fit of the binomial distribution to the data, using the Chi-square criterion of goodness of fit. To carry out this test, let $\widehat{f}\left(z; \widehat{p}\right)$ denote a numerical version of the binomial density function for $z = 0, 1, 2, \ldots, 8$ with $p = \widehat{p}$. Then, under the hypothesis that the observed sample is from a binomial distribution with index 8 and probability $\widehat{p}$, the expected number of z boys in families with 8 children would be given by

$$E(z) = M\widehat{f}(z;\widehat{p}) \tag{1.12.37}$$

for $z = 0, 1, 2, \ldots, 8$ and the Chi-square criterion of goodness of fit for the binomial distribution is

$$\mathbb{X}^2_{binom} = \sum_{z=0}^{8} \frac{(M(z) - E(z))^2}{E(z)}. \tag{1.12.38}$$

As one parameter has been estimated in this fit of the binomial distribution to the data, the number of degrees of freedom associated with $\mathbb{X}^2_{binom}$ is $9 - 1 - 1 = 7$.

One could also fit the beta-binomial distribution to the data, but in carrying out this fit, two parameters $\alpha > 0$ and $\beta > 0$ need to be estimated from the data. To estimate these parameters by the method of moments, the theoretical expectation and variance are set equal to empirical mean and variance. Let $\widehat{\alpha}$ and $\widehat{\beta}$ denote the estimates of the parameters α and β, using the method of moments. Then, $\widehat{\alpha}$ and $\widehat{\beta}$ satisfy the equations

$$8\frac{\widehat{\alpha}}{\widehat{\alpha} + \widehat{\beta}} = \overline{Z} \tag{1.12.39}$$

and

$$\frac{8\widehat{\alpha}\widehat{\beta}\left(\widehat{\alpha} + \widehat{\beta} + 8\right)}{\left(\widehat{\alpha} + \widehat{\beta}\right)^2\left(\widehat{\alpha} + \widehat{\beta} + 1\right)} = \widehat{\sigma}^2 \ . \tag{1.12.40}$$

It will be left as an exercise for the reader to solve these equations for $\widehat{\alpha}$ and $\widehat{\beta}$.

To calculate the Chi-square test criterion to fit the beta-binomial distribution to the data, let $\widehat{f}_{beta}\left(z;\widehat{\alpha},\widehat{\beta}\right)$ denote the numerical version of the beta-binomial density in (1.12.17) based on the estimates $\widehat{\alpha}$ and $\widehat{\beta}$ with $N = 8$. Then, under the hypothesis that the data are a sample from the beta-binomial distribution, the expected number of z boys in families with 8 children is

$$E(z) = M\widehat{f}_{beta}\left(z;\widehat{\alpha},\widehat{\beta}\right) \tag{1.12.41}$$

for $z = 0, 1, 2, \ldots, 8$. Given these expectations, the Chi-square criterion of goodness of fit to the data for the beta-binomial distribution would be calculated according to the formula in $(1.12.38)$.

In this case, however two parameters were estimated form the data, the number of degrees of freedom $\mathbb{X}^2$ would be 6. It is recommended that a reader use some statistical package to compute the p-values for the chi-square criterion for each of the fits of the binomial and beta-binomial distribution to the German family data. The technique used in computing these p-values is as follows. Let U_ν denote a chi-square random variable with $\nu \geq 1$ degrees of freedom. Then, given an observed value of Chi-square $\mathbb{X}$, the p-value is the numerical version of the probability

$$P\left[U_\nu > \mathbb{X}\right] \tag{1.12.42}$$

For the case of the German family data with 8 children, which distribution, the binomial or beta-binomial, yielded the best fit to the data?

Even if neither of these distributions yields a satisfactory fit to the German family data, they are both very useful as metaphors that aid our thinking about techniques to come to grips with the variation that is often observed in genetic data.

There is a generalization of the beta-binomial distribution called the Dirichlet-Mutinomial, but none of the details will be given here. The book Wilks (1962) may be consulted for a detailed account of the Dirichlet distribution, which is a multidimensional generalization of beta distribution with the property that its moments may be expressed in terms of gamma functions.

Although it has not been mentioned, an extension of the probability space $(\Omega_1, \mathcal{A}_1, P_1)$ defined in section (1.6) has been used implicitly. For every realization of the random variable $\Theta = \theta$, construct a probability $P_1(\theta, A)$ for every $A \in \mathcal{A}_1$ by using the procedure described in section (1.6). Then, the unconditional probability measure P_2 would be defined by

$$P_2(A) = \int_0^1 g_1(\theta)\, P_1(\theta, A)\, d\theta \tag{1.12.43}$$

for every $A \in \mathcal{A}_1$. In all statements regarding unconditional distribution, the probability function P_2 has been tacitly used. In subsequent chapters, related conditioning schemes will be used extensively.

Bibliography

[1] Artin, E. (1964) **The Gamma Function**. Holt, Rinehart and Winston, New York, Chicago, San Francisco, Toronto, London.

[2] Derman, C., Gleser, L. J. and Olkin, I. (1973) **A Guide to Probability and Application**. Holt, Rinehart and Winston, Inc., New York, Chicago, San Francisco, Toronto, London.

[3] Feller, W. (1968) **An Introduction to Probability Theory and Its Applications. Volume I, Third Edition**. John Wiley and Sons, New York, London, Sydney.

[4] Snustad, D. P. and Simmons, M. J. (2006) **Principles of Genetics**. John Wiley & Sons, Inc., New York, London.

[5] Wilks, S. (1962) **Mathematical Statistics**. John Wiley & Sons, Inc., New York, London.

Chapter 2

Linkage and Recombination at Multiple Loci

2.1 Introduction

Among those with a background in genetics and an interest in mathematics, a book Devlin (2000) with the title, The Math Gene, at first sight might be considered an interesting read, because of the expectation that it would contain something of interest on the genetics of mathematical ability. However, as one would find in the prologue of the book, the author is using this title as a metaphor for man's genetic predisposition for language and that mathematics is merely one aspect of language that entails the use of abstract symbols to create conceptual metaphors that lead to actions taken by a group or an individual to accomplish tasks that would not otherwise be undertaken.

An example of such group actions, which are based on mathematical metaphors underlying the laws of motion as expressed in terms of differential equations, are those that resulted in the placing of a space station in orbit around the earth so that the universe may be observed free of the distortions caused by man's activities interacting with the earth's atmosphere. Of course, fast computers to do the necessary arithmetic and rocket technology also played essential roles in such actions along with substantial amounts of public funding.

A good example of an individual's use of conceptual metaphors, based on mathematics as well as insightful observations and deductions on animals and plants, is Gregor Mendel, the founder of genetics as it was practiced in the early decades of the 20^{th} century, who studied mathematics at the University of Vienna with a goal of becoming a teacher of the subject. An elegant and insightful account of Mendel's experience at the University of Vienna and subsequent life has been provided by Bronowski (1973). For those readers with an interest in history, the text book Sinnott, Dunn and Dobzhansky (1950), a book widely read by students of genetics in the 1940's and 1950's, may be consulted for an English translation of Mendel's ground breaking paper published in 1865, reporting his experiments in hybridization of garden peas, a species which reproduces by self fertilization. One of Mendel's basic biological insights was what seems to have been his view of inheritance as a process that was governed by entities of a particulate nature that we now call genes such that an individual would either exhibit a given character or its absence.

As those with a background in mathematics are prone to do, Mendel used letters to represent the seven characters of the garden pea that he studied. To illustrate Mendel's notation, let A denote some character under study such as a variety of peas that in successive generations of self fertilization yields plants with long stems. Similarly, let a denote a variety that when self fertilized produces plants with short stems. Then, a hybrid of these two varieties was represented by the symbol Aa and the plants that bred true to long and short stems were denoted by the symbols A and a, respectively. Mendel noticed that when a population of hybrids Aa reproduced the next generations of plants by self fertilization, the characters A and a would tend to appear in a ratio of $A : a = 3 : 1$. Of course, in order to observe ratios that were close to $3 : 1$, Mendel needed to observe a rather large sample of plants that were produced by the self fertilization of a population of hybrids. The growing of such large populations of garden peas can entail a great deal of labor; consequently, a question arises as to whether the growing of large populations was an act of faith by Mendel or was it guided by some knowledge he had acquired while studying mathematics at the University of Vienna? Such knowledge could have also been acquired by reading in a library of his order, because Mendel belonged to an order of Monks that were teachers.

It is very difficult, if not impossible, to get definitive answers to such questions from the historical record, but it is tempting to speculate that Mendel may have known of mathematical theorem published by Bernoulli (1713) on a result that is now known as a Law of Large Numbers by those working in probability and statistics. In terms of Mendel's experiments with offspring of hybrid plants, let p be the actual or true frequency or probability, that a plant in a population of offspring of hybrids displays character A. Then, in a large population of size N of such plants, let N_A denote the number of plants observed with character A. The fraction, $f_A = N_A/N$, is the observed frequency of plants showing character A in such a population. Bernoulli's Law of Large Numbers states that if N is sufficiently large, then $| f_A - p |$ is very small with a probability tending to one or certainty as N becomes large. Some knowledge of this theoretical result, expressed as a theorem or in informal conversations among those with some knowledge of mathematics in his teaching order or from his experience at the University of Vienna, may have led Mendel as an act of faith in this theorem to continue his breeding experiments with peas for eight years so as to accumulate populations of sufficient size to verify his ideas empirically. For the case of the ratio $A : a = 3 : 1$, $p = 3/4$. Interestingly, even though Mendel had failed his examinations to become a teacher of mathematics at the University of Vienna, the knowledge of mathematics and its ways of thinking that he had acquired at the University seemed to have been very useful in conducting his experiments that led to a ground breaking paper in genetics. For those readers with an interest in history, the book, Uspensky (1937), may be consulted for an account of Bernoulli's proof of his Law of Large Numbers, a proof that may have been studied by some mathematicians in Mendel's time during the 19th century. One may also consult the section on a law of large numbers in chapter 1, where a proof is also presented based on Chebychev's inequality.

Just as in the 19th century, advances in genetics in the 20th and 21st centuries, entailing the sequencing of the DNA in human genome and that of other species, has led to the creation of a multitude of conceptual mathematical metaphors designed to deal with vast arrays of data that have been generated by these studies of DNA. Two areas dealing with such problems of data analysis go under the names genomics and bioinformatics. An example of a book dealing with statistical genomics is that of Liu (1998), and an example of a book dealing with bioinformatics is that of Ewens and Grant (2005).

Recently, with the initiation of the Personal Genome Project (PGP), see Church (2006) and Zhang *et al.* (2006), in the coming years it seems likely that large data sets, containing the sequenced genomes of many people, will become available in data bases maintained by such institutions as the National Institutes of Health (NIH). Furthermore, the likelihood that such data bases will become available within five to ten years has been increased with the posting of a 10 million dollar prize, sponsored by the X-Prize Foundation, to the first company that can sequence the genomes of 100 diverse people in ten days. Moreover, with continued advances in technology for the sequencing of entire genomes, it also seems likely that this technology will be used to sequence the genomes of animals used in basic science, such as mice and Drosophila, as well as plants of not only scientific interest but also of economic value. A personal account of the sequencing of the human and Drosophila genomes has been given in the interesting book, Venter (2007). Such large data bases will very likely give rise to the possibility of studying genetic recombination at the molecular level, a process that has been thought to be significant in producing genetic variability upon which natural and artificial selection has acted during the evolution of wild and domestic species of animals and plants as well those disease organisms causing their diseases. In particular, a knowledge of recombination among DNA markers could expedite the task of locating genes affecting human diseases that may be linked to DNA markers.

Accordingly, the purpose of this chapter is to develop a set of mathematical and statistical metaphors that will be helpful in organizing and analyzing data from sequenced genomes of relatives, particularly parents and their offspring, as well as theoretical calculations in theoretical population genetics. Such data sets will contain observed patterns of genetic recombination in offspring generations as an output of the process of meiosis in parental generations of diploid species. As currently planned, these data bases will also include information on phenomes or phenotypes of individuals so that some assessments may be made regarding the effects of genetic recombination on genotypes and their phenotypic expressions. As attention will be focused on patterns of recombination of maternal and paternal DNA, it should be stated at the outset that no attempt will be made in this chapter to model the process of meiosis at the molecular level, but if a reader is interested in models of meiosis at the Mendelian and cellular levels, the monograph of Thompson (2000) may be consulted. Two other interesting papers, the Models of the Chiasma processes presented in Zhao and Speed (1996) and Browning (2000), may also be consulted. With a view towards furthering the development of mathematical and statistical metaphors to deal with ideas about the evolution of recombination in humans and other species, it is recommended that the review paper Coop *et al.* (2007) be consulted.

2.2 Some Thoughts on Constructing Databases of DNA Markers From Sequenced Genomes of Relatives

When constructing mathematical models within the framework of Mendelian genetics, the terms genes, locus on a chromosome with multiple alleles at that locus have useful operational meanings, and, when linkage models concerning more than one locus are considered, one speaks of models with many loci with multiple alleles at each locus. However, when models using these terms are considered at the molecular level, such terms as multiple loci with many alleles have less clear operational meanings than in the abstractions of mathematical Mendelian genetics. It is beyond the scope of this chapter to attempt to give an account of the meaning of such terms at the molecular level, but it is recommended that books, such as Strachan and Read (2004), be consulted. The reading of such books can be a formidable task for non-experts in molecular genetics, but even a perusal of the table on contents and exploratory reading of selected sections of a book can be helpful in gleaning meanings for the terms locus or multiple loci with many alleles at the molecular level.

As a start in a search for examples for meanings of such terms at the molecular level, it is helpful to enter the title, Major Histocompatibility Complex (MHC), into a search engine for the world wide web. MHC is a large genomic or gene family found in most vertebrates. In humans, for example, the MHC region is a 3.6 Mb (3,600,000 base pairs) region on chromosome 6 and contains an estimated 140 genes between the flanking markers with the labels MOG and COL11A2, see Wikipedia, the free encyclopedia. Within the MHC region there are the human leukocyte antigen (HLA) genes. Among the so-called classical HLA genes that have been studied are the HLA-A, HLA-B, and HLA-C loci that have multiple alleles. For example, one of the alleles at the HLA-B locus, HLA-B-27, is found in many people who present clinical symptoms of ankylosing spodylitus, an autoimmune disease and a type of arthritis in which some vertebrae become fused when cartilage turns to bony tissue. There are also six HLA-D genes characterized by haplotypes with of tens of thousands of base pairs.

In humans, the MHC region is partitioned into three classes labeled I, II, and III. The A, B, and C HLA genes belong to class I and the D genes belong to class II. In a recent paper, Raymond, C. K. *et al.* (2005) studied variations in haplotypes by resequencing 20 haplotypes across about -100,000 base pairs (-100-kbp) which spans the HLA-DQA1, -DQB1 and -DRB1 genes. As a result of this study, the authors have provided a detailed tentative view of the way in which the genome structure of these loci has been shaped by evolutionary forces involving the interplay of selection, gene-gene interaction and recombination. For those readers that are interested in finding explicit examples of some of the techniques discussed in books on bioinformatics, such as problems in aligning sequences of DNA, it is helpful and useful to read and study this paper in detail.

The paper by Raymond and his colleagues is but one example of the study of genomic diversity in the human MHC. For example, see the Wikipedia article, it has been estimated that the HLA_A, HLA-B and HLA-DRB1 loci have roughly 250, 500 and 300 known alleles, respectively. This degree of polymorphism is exceptional for the human genome and has prompted populations geneticists to ask questions as to what evolutionary processes may have been acting to bring this observed high degree of polymorphism into existence. In this connection, it has been suggested that smells, olfaction, may have played a role in the selection of mates so that females and males are more likely to prefer mates with different MHC genes. In multiple allelic systems, this type of mating preferences suggests that it is likely that both parents in a mating may be heterozygous at two or more loci. It is known that MHC genes make molecules that enable the immune system to recognize foreign bodies and destroy them. This suggests that natural selection would favor a mating system that would increase the diversity of genes in the parents. It seems plausible that the more diverse the genes of the parents, the greater will be the ability of their offspring to survive attacks by diseases organisms and other adverse environmental challenges. For readers interested in more details on the subject just mentioned, it is recommended that the references in Raymond, C. R. *et al.* (2005) as well as those in the Wikipedia article be consulted.

As an illustrative example, suppose an investigator wanted to investigate whether the of sets of nucleotides among these the 9 HLA loci may be implicated in some autoimmune disease under study. Moreover, suppose an investigator wanted to investigate whether the effects of genetic recombination of parental genes may have on the expression of the disease in their children. An investigator may also be interested in investigating the effects of genetic recombination on the regulatory functions of DNA. Other sets of linked genetic markers could also be considered in searches of regions of DNA among these markers that are associated with observed diseases. In such studies, it would be of interest to search for micro-structure in the DNA separating sets of linked loci. One of the most common micro structures in the human genome are $SNP's$, single nucleotide polymorphisms, characterized by a single base substitutions, see Boxes 7.2 and 13.1 in Strachan and Read (2004). It has been estimated, for example, that the number of $SNP's$ in the human genome is greater than 4×10^6.

Other micro structures of DNA include VNTR (variable number of tandem repeats) polymorphisms. Among these polymorphisms are the microsatellites VNTR, which are frequently less that 100 base pairs long with repeat units 1 to 4 nucleotides long. Further examples are discussed in Strachan and Read (2004).

As a first step of a thought experiment on creating data bases from data bases of sequenced genomes of relatives, suppose $N \geq 2$ loci of interest have been identified in the data. Each locus, for example, may consist of thousands or millions of base pairs, a single nucleotide for the case of $SNP'S$ or hundreds of base pairs for the case of VNTR polymorphisms. Suppose, for the sake of simplicity, that each locus in this set is on the same chromosome or linkage group and that sequenced data from the parents and their children and perhaps grandchildren are available at each locus under consideration in the data base.

As a next step of a thought experiment in constructing a data base of sequenced parental and offspring DNA, suppose data are available on $n > 1$ couples whose DNA has been sequenced and suppose that the *i-th* couple has $n_i \geq 1$ offspring or children whose DNA has also been sequenced. Then, the total number of observations on the children available for the study of genetic recombination would be

$$m = \sum_{i=1}^{n} n_i. \tag{2.2.1}$$

In the next section, some concepts will be discussed that will be helpful in selecting those matings among the n couples and their n_i children that will be informative with respect to the study of genetic recombination and its effects on the expression of disease in their children and other descendants.

It is of interest to note that investigators have already assembled rather large data bases of the DNA of relatives. For example, in a very interesting paper Coop *et al.* (2008) have assembled a data base consisting of 1650-person pedigree. From this pedigree it was possible to infer recombination events in 364 female and 364 male gametes, which revealed extensive variation in fine-scale recombination patterns. In summarizing the data, the authors focused on statistical summaries such as histograms of the number of recombination events in interval sizes expressed in kilo bases and variation in recombination rates among individuals. Such summaries point to the conclusion that meiosis is a complex stochastic process that warrants further attention by those interested in developing stochastic models of evolutionary processes at the molecular level.

2.3 Examples of Informative Matings for the Case of Two Loci

In this section, some examples of couples, matings, will be given such that evidence of genetic recombination in the parents can be detected among their children. Such matings will be called informative with respect to genetic recombination. Illustrative examples will be presented for the case of two loci with two or more alleles at each locus under study, because even the description of informative matings with respect to many loci with many alleles is a formidable task that will not be attempted in this section. However, a system of notation for representing haplotypes and genotypes will be described for cases of two loci with two alleles at each locus that can easily be extended to cases multiple loci with multiple alleles at each locus.

To help fix ideas, suppose the DNA on the chromosome in a large sample of couples has been sequenced so that the average number of base pairs making up each locus in this sample of individuals is known. Moreover, suppose the average number of base pairs separating the two loci is also known and that two distinct haplotypes at each locus may be recognized at the molecular level. Let $(a_{1i} \mid i = 1, 2)$ denote the set of recognizable haplotypes (alleles) at locus 1 and let $(a_{2j} \mid j = 1, 2)$ denote the set of recognizable haplotypes (alleles) at locus 2. Before constructing arrays of haplotypes and genotypes with respect to two loci, it will be helpful to define some index sets. Let $\mathbb{I} = (i \mid i = 1, 2)$ denote the index set for two alleles at each of two loci, and let

$$\mathbb{I}^2 = ((i,j) \mid i \in \mathbb{I}\ ;\ j \in \mathbb{I}) \tag{2.3.1}$$

denote a product set of indices for two loci.

Then, the set or array of recognizable haplotypes with respect to two loci may be represented in the product form

$$\left(a_{1i}a_{2j} \mid (i,j) \in \mathbb{I}^2\right). \tag{2.3.2}$$

Observe that this product array will contain $2 \times 2 = 4$ symbols, representing four recognizable haplotypes with respect to two loci. If one thinks of this product set of haplotypes with respect to two loci as a set of possible gametes arising from the cell division process of meiosis, then the set of possible genotypes with respect to the set of haplotypes at the two loci may be represented as the quotient set

$$\left(\frac{a_{1i}a_{2j}}{a_{1k}a_{2l}} \mid \left(i,j\right) \in \mathbb{I}^2; \left(k,l\right) \in \mathbb{I}^2\right) \tag{2.3.3}$$

Note that this quotient set contains $2^4 = 16$ genotypes. When representing genotypes in this quotient form, the symbol in the upper level of the quotient may be thought of as the gamete contributed by the female parent of an individual and the lower level symbol may be thought of as the gamete contributed by the male parent. The quotient representation of a genotype is particularly helpful when one is thinking about cross overs between two loci that may occur during meiosis.

In order to simplify the writing of the symbols, particularly in the form of matrices or tables, it will be useful to let the subscripts represent the haplotypes or gametes under consideration by setting up the correspondence

$$a_{1i}a_{2j} \leftrightarrow ij, \tag{2.3.4}$$

with the understanding that the symbol i stands for the haplotype a_{1i} and the symbol j stands for the haplotype a_{2j}. Given this convention, it follows that an arbitrary genotype may be represented by the more succinct notation as the quotient

$$\frac{a_{1i}a_{2j}}{a_{1k}a_{2l}} \leftrightarrow \frac{ij}{jl} \leftrightarrow \left(ij, kl\right). \tag{2.3.5}$$

or in the symbols on the right. The symbol on the far right is particularly useful when displaying genotypes in table or matrix form.

Given this more succinct notation, the array of four haplotypes with respect to two loci may be represented as the ordered array

$$\left(11, 12, 21, 22\right). \tag{2.3.6}$$

Then, given this ordered array of gametes expressed symbolically as haplotypes, the array of 16 genotypes with respect to these haplotypes may be represented as a set of ordered pairs in the 4×4 table

$(11,11)$	$(11,12)$	$(11,21)$	$(11,22)$
$(12,11)$	$(12,12)$	$(12,21)$	$(12,22)$
$(21,11)$	$(21,12)$	$(21,21)$	$(21,22)$
$(22,11)$	$(22,12)$	$(22,21)$	$(22,22)$

. (2.3.7)

On the principal diagonal of this table, the cells making up the left upper cell to right lower cell, contain genotypes that are all homozygous. Hence, they would not be informative for studying recombination events in meiosis between two loci. But, on the diagonal making up those cells of the table from the lower left cell to the upper cell on the right, all genotypes are heterozygous, and, therefore, in suitable matings with these genotypes recombination events in meiosis could be observed in the offspring or children.

If an investigator observed the genotypes of a set of parental couples, then, in the absence of suitable DNA markers, there would be no way of determining whether a haplotype originated from the female or male parent. Therefore, only two kinds of heterozygous genotypes at two loci could be recognized in a set of individuals or couples. For example, the genotypes expressed in quotient form

$$\frac{22}{11} \approx \frac{11}{22} \tag{2.3.8}$$

would be equivalent, because they have the same haplotypes and would, therefore, not be distinguishable. With respect to recombination during meiosis, in the genotype on the right, the haplotypes 11 and 22 would represent nonrecombinant events during meiosis, and, the haplotypes 12 and 21 would represent recombination events.

Similarly, the two genotypes

$$\frac{21}{12} \approx \frac{12}{21} \tag{2.3.9}$$

would have the same haplotypes so that in this sense they would be equivalent. However, with respect to the heterozygous genotype on the right, the haplotypes types 12 and 21 would represent non-recombinant events during meiosis, while the haplotypes 11 and 22 would represent recombinant events. As these two examples show, whether haplotypes represent non-recombinant events or recombinants events in meiosis depends on the phase of the haplotypes in the parents. Some authors refer to the first example as the coupling phase, while the second examples is sometimes referred to as the repulsion phase.

Now suppose both the female and males in couple have been sequenced and it has been found that the genotype of the female is $(11, 22)$ in coupling phase and that of the male is $(11, 11)$ so that he is homozygous at both loci. Such homozygous individuals would produce only gametes of the form 11. Then, it follows that the children of this mating, symbolized by

$$\frac{11}{22} \otimes \frac{11}{11}, \tag{2.3.10}$$

would have genotypes belonging to the set

$$\left(\frac{11}{11},\frac{12}{11},\frac{21}{11},\frac{22}{11}\right). \tag{2.3.11}$$

The haplotypes 12 and 21 represent recombination events during meiosis in the female parent; whereas, the haplotypes 11 and 22 indicate that no recombination occurred in the female parent. Actually, in a genotypes of the form $(11, 11)$ it may not be possible to determine which haplotype was contributed by the female parent. However, in such homozygous individuals from matings of the type under consideration, it is certain that one haplotype resulted from a nonrecombination event in the female parent. The type of matings under consideration are called informative with respect to recognizing recombination events, because recombinant haplotypes of the female can be recognized unambiguously in the children of such matings. Furthermore, if the male was of genotype $(22, 22)$ in the example under consideration, then the mating would also be informative with respect to recombination events in the female. Observe matings of this type would also be informative with respect of recombination events in the male, if the genotypes of the female and male were interchanged.

When sequenced data are available for couples and their offspring, one may find examples of matings that are informative with respect to recombination in both female and male parents but among the offspring of such matings, it may not be possible to determine whether the recombination event occurred in the female or male parent. An example of such of mating is

$$\frac{11}{22}\otimes\frac{11}{22}. \tag{2.3.12}$$

Because recombination gametes in both parents have the same haplotypes, in the absence of other DNA markers, it would not be possible to distinguish among the haplotypes in the offspring whether they came from the female or male parent.

In systems with multiple alleles systems and linked genes at several loci, such as the HLA systems in humans, it will, in principle, be possible to find matings in which recombination events in both the female and male parents can be detected in their children. Consider, for example, the mating

$$\frac{12}{34}\otimes\frac{56}{78} \tag{2.3.13}$$

in which each parent is heterozygous with respect to four alleles. If it is indeed the case that natural selection favors those matings involving diverse alleles, then one would expect to find such matings in greater frequency than if the choice of mates were purely random. In females of such matings, the gametes are the set of haplotypes

$$(12, 14, 32, 34)\,. \tag{2.3.14}$$

In this set the haplotypes 12 and 34 represent non-recombinant events and the haplotypes 14 and 32 represent recombinant events in meiosis. For males, the gametes are the set of haplotypes

$$(56, 58, 76, 78)\,, \tag{2.3.15}$$

but it will be left as an exercise for the reader to determine which haplotypes are recombinants or non-recombinants. If these orderings of the gametes for females and male parents are used, then the possible genotypes of the children of such matings can be represented in the 4 × 4 table of ordered pairs

$(12, 56)$	$(12, 58)$	$(12, 76)$	$(12, 78)$
$(14, 56)$	$(14, 58)$	$(14, 76)$	$(14, 78)$
$(32, 56)$	$(32, 58)$	$(32, 76)$	$(32, 78)$
$(34, 56)$	$(34, 58)$	$(34, 76)$	$(34, 78)$

. (2.3.16)

From an inspection of this table, it can be seen that in this example involving 8 alleles at two loci, it would be possible to classify each haplotype in the children as to whether it is a non-recombinant or recombinant with respect to either the mother or father of the a child. Therefore, children of matings of the type under consideration would be very informative with respect to genetic recombination in both the female and male parents. In systems with multiple loci with many alleles at each locus, such as the HLA systems in humans. One would expect to find matings with a high diversity of alleles in both the females and males. For those readers that may be interested in further examples of types of matings that will be informative with respect to recombination in their children, the book Liu (1998) may be consulted.

It would be useful to construct examples of informative matings with respect to detecting recombination events in the children of such matings for the cases of three loci, but no attempts will be made in this section to present these more complicated examples. One of the very challenging problems in connection with working multiple loci systems with multiple alleles is to set down a notational systems that would be amenable to the development of software that would detect informative matings with respect of genetic recombination in large data sets containing information in the form of sequenced DNA of parents and their children.

2.4 General Case of Two Linked Loci

In a diploid individual a gene at a particular locus is either of maternal or paternal origin. Therefore, to develop a metaphor for a recombination probability in this case, it will suffice to indicate only whether a gene is of maternal or paternal origin, because in the study of recombination at either the Mendelian or molecular level, the focus of attention is whether a gene in an offspring is either of maternal of paternal origin. A gene of maternal origin will be denoted by the Boolean symbol 0, and a gene of paternal origin will be denoted by the symbol 1. Given this notation, an arbitrary genotype of diploid individual with respect to two linked loci may be represented in the form

$$\frac{00}{11}, \tag{2.4.1}$$

symbolizing the contribution from each parent. An individual of this genotype may in turn produce four types of gametes; namely $00, 01, 10$, and 11. Gametes 00 and 11 represent copies of the maternal and paternal gametes respectively; while 01 and 10 are recombination type gametes containing both maternal and paternal genes. In particular cases, the generic gametic symbols could be identified with particular haplotypes when the focus of attention is at the molecular level, or with phenotypes when the discussion is at the Mendelian level. It should also be noted that the Boolean notation under consideration is not phase dependent with respect to some specific genes at two loci to which attention is being directed, because in terms of Boolean indicators only parental origins of genes are under consideration.

The four types of gametes will be produced by a given genotype with certain probabilities depending on the probability of recombination. Let $\gamma(00)$, $\gamma(01)$, $\gamma(10)$, and $\gamma(11)$ denote the probabilities an arbitrary genotype produces gametes 00, 01, 10, and 11, respectively. In the two loci case, the linkage (gametic) distribution is the set

$$(\gamma(00), \gamma(01), \gamma(10), \gamma(11)) \tag{2.4.2}$$

of non-negative numbers whose sum is one. Due to the complementary nature of the meiotic mechanism, i.e., gametes are almost always produced in pairs, the mechanism of meiosis will be called balanced if the equations

$$\begin{aligned}\gamma(00) &= \gamma(11)\\ \gamma(10) &= \gamma(01)\end{aligned} \tag{2.4.3}$$

are satisfied.

If we let ρ be the probability of recombination, then it follows that

$$\gamma(10) + \gamma(01) = \rho \tag{2.4.4}$$

and

$$\gamma(00) + \gamma(11) = 1 - \rho. \tag{2.4.5}$$

But, if the mechanism of meiosis is balanced, then

$$\gamma(01) = \gamma(10) = \frac{1}{2}\rho \tag{2.4.6}$$

and

$$\gamma(00) = \gamma(11) = \frac{1}{2}\left(1 - \rho\right). \tag{2.4.7}$$

When an objective of an investigation is to map the locations of two loci on the same chromosome expressed in terms of centimorgens (cM), then the pertinent values of ρ will satisfy the condition $0 \leq \rho \leq \frac{1}{2}$. The case of random assortment occurs when $\rho = \frac{1}{2}$ so that the probability of each type of gamete is 1/4. In principle, however, ρ may be any value such that $0 \leq \rho \leq 1$.

A problem of considerable theoretical importance in assessing the effects of linkage in populations is that of defining recombination probabilities and finding a relation between these recombination probabilities and the gametic distribution, when the number of loci under consideration is greater than four. An interesting step towards a solution of this problem was made by Schnell (1961), who observed that if one makes the transformation

$$\rho = \frac{1}{2}\left(1 - \lambda\right), \tag{2.4.8}$$

or equivalently

$$\lambda = 1 - 2\rho, \tag{2.4.9}$$

then a certain orthogonality is introduced which leads to an extension to cases of an arbitrary number of linked loci. Observe that as ρ varies over the interval $[0, 1]$, the parameter λ varies over the interval $[-1, 1]$.

By using these equations, it can be seen that

$$\begin{aligned} \gamma(00) &= \frac{1}{4}\left(1 + \lambda\right) \\ \gamma(10) &= \frac{1}{4}\left(1 - \lambda\right). \end{aligned} \tag{2.4.10}$$

These equations can perhaps be most easily comprehended in vector-matrix notation. Let

$$\begin{aligned}\boldsymbol{\gamma}^T &= (\gamma(00), \gamma(10)) \\ \boldsymbol{\lambda}^T &= (1, \lambda) \end{aligned} \tag{2.4.11}$$

be 1×2 vectors, where the superscript T stands for transpose of a vector or matrix, and let

$$\mathbf{A}_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \tag{2.4.12}$$

denote a 2×2 matrix. Observe that this matrix is orthogonal. Then, in vector-matrix notation the above equation may be written in the form

$$\boldsymbol{\gamma} = \frac{1}{4}\mathbf{A}_2\boldsymbol{\lambda}. \tag{2.4.13}$$

From now on the superscript T on a vector will stand for the transpose of a vector or matrix.

The following observations are very helpful in finding an extension to the case of an arbitrary number of linked loci. Firstly,

$$\mathbf{A}_2^T = \mathbf{A}_2 \tag{2.4.14}$$

so that $\mathbf{A}_2$ transpose is $\mathbf{A}_2$. When a square matrix remains invariant under the operation of transposition, it is said to be symmetric. Secondly, observe that

$$\mathbf{A}_2^T\mathbf{A}_2 = \mathbf{A}_2^2 = 2\mathbf{I}_2, \tag{2.4.15}$$

where $\mathbf{I}_2$ is a 2×2 identity matrix. This equation expresses an orthogonality condition which, as we shall see, may be generalized to an arbitrary number of loci greater than two. Thirdly, if the symbols 00 and 10 are regarded as the vectors $\boldsymbol{\xi}_1^T = (0, 0)$ and $\boldsymbol{\xi}_2^T = (1, 0)$, then the matrix $\mathbf{A}_2$ may be represented in the form

$$\mathbf{A}_2 = (a_{ej}) = \left((-1)^{\boldsymbol{\xi}_i^T\boldsymbol{\xi}_j}\right), \tag{2.4.16}$$

for $i, j = 1, 2$. Lastly, we see the column vector $\boldsymbol{\lambda}$ may be represented in the form

$$\boldsymbol{\lambda} = 2\mathbf{A}_2\boldsymbol{\gamma}. \tag{2.4.17}$$

This equation could be taken as a definition of the column vector $\boldsymbol{\lambda}$. In the next section, these results will be extended to the case of three linked loci.

Before considering the case of three loci however, it will be of interest to discuss briefly an illustrative procedure for estimating the recombination probability ρ based on the example in the foregoing section on matings involving systems with multiples alleles such that it was possible to distinguish in the haplotypes of the offspring whether a recombination event had occurred in female or male parent of a child. With regard to this example, let $n\,(ij, kl)$ be the total number of children in a sample with haplotypes (ij, kl), where haplotypes ij were those from the mother and kl those from the father. Both females and male children may be included in these counts. Given these counts, the number of recombinant haplotypes from the mother of type 14 would be sum

$$n\,(14) = \sum_{kl} n\,(14.kl)\,. \tag{2.4.18}$$

Similarly, the total number of recombinant haplotypes from the mother of type 32 would be

$$n\,(32) = \sum_{kl} n\,(32.kl)\,. \tag{2.4.19}$$

In large samples of matings of the types under consideration, it would be expected that the numbers $n\,(14)$ and $n\,(32)$ would be nearly equal, because of the balanced nature of meiosis. In small samples however, these numbers may differ. Let n_{tot} denote the total number of children in the sample. Then,

$$\widehat{\rho}_f = \frac{n\,(14) + n\,(32)}{n_{tot}} \tag{2.4.20}$$

would be an estimate of the recombination probabilities in females. For matings of the type under consideration, a recombination probability in males could be estimated in a similar way but the formal details will be omitted. Briefly, when several mating types have been observed in the data such that it was possible to distinguish in the children whether a recombination type of haploid was contributed by the mother or father in the matings, it would be necessary to pool counts over types of matings to get estimates of the probabilities of recombination in females and males.

2.5 General Case of Three Linked Loci

A question that naturally arises is whether the above scheme may be generalized to any arbitrary number of loci greater than two. As we shall see such a generalization is possible, but the manner in which the scheme may be generalized will not become clear until the three loci case is considered. In the three loci case, an arbitrary diploid genotype may be represented in the form

$$\frac{000}{111}, \tag{2.5.1}$$

where as before the zeros and ones represent genes contributed by maternal and paternal parents, respectively. This genotype is capable of generating $2^3 = 8$ types of gametes containing various combinations of maternal and paternal genes.

The set of these eight types of gametes will be represented in the form

$$\mathbb{G} = (000, 100, 010, 110, 111, 011, 101, 001)\,, \tag{2.5.2}$$

and let the vector

$$(\gamma(\xi) \mid \xi \in \mathbb{G}) \tag{2.5.3}$$

denote the linkage distribution, *i.e.*, the set of non-negative numbers, whose sum is one, giving the probability that each type of gamete is produced by the meiotic process. Because the meiotic mechanism is assumed to be balanced, gametes are produced in pairs so that following symmetry or complementary conditions

$$\begin{aligned}\gamma(000) &= \gamma(111)\\ \gamma(100) &= \gamma(011)\\ \gamma(010) &= \gamma(101)\\ \gamma(110) &= \gamma(001)\end{aligned} \tag{2.5.4}$$

will be assumed. From these symmetry conditions, it follows that it will be sufficient to consider only four probabilities from the linkage distribution in setting up a correspondence with a set of recombination probabilities.

In the three loci case, a recombination probability may be associated with each pair of loci; namely, the pairs $(1, 2), (1, 3)$ and $(2, 3)$. Let ρ_{12}, ρ_{13}, and ρ_{23}, be the probability of recombination between the respective pairs of loci. With each recombination probability, we may associate a lambda parameter denoted by λ_{12}, λ_{13} and λ_{23}. From the definitions of ρ_{12}, ρ_{13}, and ρ_{23}, it follows that

$$\begin{aligned}\rho_{12} &= 2\,(\gamma(100) + \gamma(010))\\ \rho_{13} &= 2\,(\gamma(100) + \gamma(110))\\ \rho_{23} &= 2\,(\gamma(010) + \gamma(110))\\ 1 - \rho_{12} &= 2\,(\gamma(000) + \gamma(110))\,.\end{aligned} \tag{2.5.5}$$

Observe that only four gametic probabilities appear on the right so it may be possible to solve four simultaneous linear equations. Also observe that the symmetry conditions were used in choosing the four gametic probabilities on the right.

By substituting the $\lambda's$ for the $\rho's$ in these equations, it can be shown that

$$\begin{aligned}\frac{1}{4}(1-\lambda_{12}) &= \gamma(100)+\gamma(010)\\ \frac{1}{4}(1-\lambda_{13}) &= \gamma(100)+\gamma(110)\\ \frac{1}{4}(1-\lambda_{23}) &= \gamma(010)+\gamma(110)\\ \frac{1}{4}(1+\lambda_{12}) &= \gamma(000)+\gamma(110). \qquad (2.5.6)\end{aligned}$$

By solving these four equations for $\gamma(000), \gamma(100), \gamma(010)$ and $\gamma(110)$, it can be seen that

$$\begin{aligned}\gamma(000) &= \frac{1}{8}(1+\lambda_{13}+\lambda_{23}+\lambda_{12})\\ \gamma(100) &= \frac{1}{8}(1-\lambda_{13}+\lambda_{23}-\lambda_{12})\\ \gamma(010) &= \frac{1}{8}(1+\lambda_{13}-\lambda_{23}-\lambda_{12})\\ \gamma(110) &= \frac{1}{8}(1-\lambda_{13}-\lambda_{23}+\lambda_{12}). \qquad (2.5.7)\end{aligned}$$

Just as in the two loci case, these equations may be most easily comprehended if they are cast in vector-matrix notation. Let

$$\boldsymbol{\gamma}^T = (\gamma(000), \gamma(100), \gamma(010), \gamma(110)) \qquad (2.5.8)$$

and

$$\boldsymbol{\lambda}^T = (1, \lambda_{13}, \lambda_{23}, \lambda_{12}) \qquad (2.5.9)$$

denote row vectors, and let

$$\mathbf{A}_3 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}. \qquad (2.5.10)$$

denote a 4 × 4 matrix. Then the equations may be written in the compact form

$$\boldsymbol{\gamma} = \frac{1}{2^3}\mathbf{A}_3\boldsymbol{\lambda}. \qquad (2.5.11)$$

It should also be noted that the matrix $\mathbf{A}_2$ is related to the matrix $\mathbf{A}_3$ by the simple recursion formula

$$\mathbf{A}_3 = \begin{bmatrix} \mathbf{A}_2 & \mathbf{A}_2 \\ \mathbf{A}_2 & -\mathbf{A}_2 \end{bmatrix}. \tag{2.5.12}$$

From this recursive equation, it can be seen that

$$\mathbf{A}_3^T = \begin{bmatrix} \mathbf{A}_2^T & \mathbf{A}_2^T \\ \mathbf{A}_2^T & -\mathbf{A}_2^T \end{bmatrix} = \begin{bmatrix} \mathbf{A}_2 & \mathbf{A}_2 \\ \mathbf{A}_2 & -\mathbf{A}_2 \end{bmatrix} = \mathbf{A}_3 \tag{2.5.13}$$

because $\mathbf{A}_2$ is symmetric. Furthermore, from this equation, it follows that

$$\begin{aligned} \mathbf{A}_3^T \mathbf{A}_3 = \mathbf{A}_3^2 = \begin{bmatrix} 2\mathbf{A}_2^2 & \mathbf{0} \\ \mathbf{0} & 2\mathbf{A}_2^2 \end{bmatrix} = 2 \begin{bmatrix} 2\mathbf{I}_2 & \mathbf{0} \\ \mathbf{0} & 2\mathbf{I}_2 \end{bmatrix} \\ = 2^2 \mathbf{I}_4 \end{aligned} \tag{2.5.14}$$

where $\mathbf{0}$ is a 2 × 2 zero matrix and $\mathbf{I}_4$ is an identity matrix of order 4. We thus see that orthogonality condition carries over to the three loci case. Moreover, let $\boldsymbol{\xi}_1^T = (0,0,0), \boldsymbol{\xi}_2^T = (1,0,0), \boldsymbol{\xi}_3^T = (0,1,0)$ and $\boldsymbol{\xi}_4^T = (1,1,0)$ denote row vectors. Then, by inspection, it can be seen that

$$\mathbf{A}_3 = \left(a_{ij} = (-1)^{\boldsymbol{\xi}_i^T \boldsymbol{\xi}_j} \right) \tag{2.5.15}$$

for all $i, j = 1, 2, 3, 4$. Just as in the case of two loci, it also follows, by using the above results and solving for the vector $\boldsymbol{\lambda}$, that the equation

$$\boldsymbol{\lambda} \; = 2\mathbf{A}_3 \boldsymbol{\gamma}, \tag{2.5.16}$$

expresses the vector $\boldsymbol{\lambda}$ as a linear function of the vector $\boldsymbol{\gamma}$, given the matrix $\mathbf{A}_3$.

At this juncture it is important to note that the ordering of the gametic symbols

$$(000, 100, 010, 110) \tag{2.5.17}$$

played a basic role in extending the observations made in the two loci case to the case of three loci. Furthermore, the linear relation (2.5.16), connecting the vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$, suggests that to extend the results for the case of three loci to cases of four or more loci, it would be prudent to simplify the notation by abandoning subscripts on the λ and ρ parameters and replacing them by a function notation of the form $\lambda(\xi)$ and $\rho(\xi)$, where ξ is an arbitrary gametic symbol containing $0's$ and $1's$.

Given this function notation, the matrix $\mathbf{A}_3$ and the ordering of the elements of the gametic vector $\boldsymbol{\gamma}$, equation (1) leads to an automatic ordering of the elements of the vector $\boldsymbol{\lambda}$. This observation suggests that extensions of equation (1) to cases of four or more loci could be used to define the vector $\boldsymbol{\lambda}$ for an arbitrary number of loci. Then, given any $\xi \in \mathbb{G}$, a corresponding recombination probability may be determined by the equation

$$\lambda(\xi) = 1 - 2\rho(\xi). \tag{2.5.18}$$

By way of an illustrative example, suppose $\xi = 000$, indicating that in gametes of this type there was no genetic recombination. For the case of three loci under consideration, it was observed that $\lambda(\xi) = 1$, which implies $\rho(\xi) = 0$, indicating with gametes of type $\xi = 000$ recombination occurs with probability 0, which is consistent with our intuition.

It would be of interest to present an example of data such that the probabilities of recombination for the case of three linked loci could be estimated, but no attempt to construct such an example will be undertaken. For readers who are interested in three point test for linkage, the book by Liu may be consulted.

2.6 General Case of Four or More Linked Loci

As suggested in the previous section, the key to a solution of our problem of extending the cases of two and three linked loci to cases of four or more loci lies in a procedure to order the elements in the set $\mathbb{G}$ of gametic symbols or recombination patterns. As a first step in defining this ordering procedure, consider the case of deriving the set $\mathbb{G}_3$ for the case of three linked loci from the set

$$\mathbb{G}_2 = (00, 10) \tag{2.6.1}$$

for the case of two linked loci. Starting with the symbols $00, 10$ in the two loci case, the complementary set of gametes in the set $\mathbb{G}_3$ may be generated according to the following procedure. From the symbols 00 and 10 construct two symbols by adding a zero in the third position to obtain symbols the symbols 000 and 100. Then construct two additional symbols from 00 and 10 by adding a one in the second position and zero in the third position to obtain 010 and 110. This procedure thus produces the ordered set

$$\mathbb{G}_3 = (000, 100, 010, 110) \tag{2.6.2}$$

for the case of three loci. Observe that this was the ordering used for extending the two loci case to the three loci case in section 2. As before this set of symbols may be regarded as vectors by letting $\boldsymbol{\xi}_1^T = (0,0,0)$, $\boldsymbol{\xi}_2^T = (1,0,0)$, $\boldsymbol{\xi}_3^T = (0,1,0)$, and $\boldsymbol{\xi}_4^T = (1,1,0)$.

Before proving that the above results will extend to the general case of N linked loci, let us pause to fix ideas and to illustrate the procedure outlined above for case $N = 4$. If $N = 4$, then $N - 1 = 3$ and the set of gametic symbols to be considered is $(000, 100, 010, 110)$. The set of four symbols $(0000, 1000, 0100, 1100)$ was obtained by adding a zero in the 4-*th* position, and then the set of four symbols $(0010, 1010, 0110, 1110)$ was obtained by inserting a one in the third position and a zero in the fourth position. The set of ordered gametic symbols for the four loci case is, therefore,

$$\mathbb{G}_4 = (0000, 1000, 0100, 1100, 0010, 1010, 0110, 1110). \tag{2.6.3}$$

This set of symbols may be interpreted as a set of eight vectors

$$(\boldsymbol{\xi}_i \mid i = 1, 2, \ldots, 8) \tag{2.6.4}$$

ordered in the way they appear from left to right.

As a last step before returning to the mathematics of the model, it is appropriate to give a biological interpretation of $\rho\,(1110)$, the recombination parameter corresponding to the lambda parameter $\lambda\,(1110)$. Formally,

$$\rho\,(1110) = \frac{1 - \lambda\,(1110)}{2}. \tag{2.6.5}$$

Observe that the type of gamete corresponding to these parameters is (1110) and its complement is (0001). In gametes of this type, the maternal and paternal genes have been preserved in the process of meiosis at loci $1, 2$ and 3. But, there has been a crossover between loci 3 and 4 in that maternal and paternal genes have been interchanged. In other words, in this product of meiosis, loci $1, 2$ and 3 behaved as a unit and there was a cross over between this unit and locus 4. Stated in terms of the number of crossovers, one can see that there has been an even number of crossover between loci 1 and 2 as well as between loci 2 and 3 and odd number of crossovers between the loci 3 and 4.

To proceed to the general case, suppose we have arrived at the case of $N - 1$ loci by repeating the procedure outlined above. At the $N - 1$ loci stage, we will have generated an ordered set

$$\mathbb{G}_{N-1} = (\boldsymbol{\xi}_i \mid i = 1, 2, \ldots, 2^{N-2}) \tag{2.6.6}$$

of 2^{N-2} gametic symbols containing zeros and ones. Each of these symbols may also be interpreted as a $N-1$ dimensional vector and, moreover, the set of vectors is ordered according to the order in which the gametic symbols were generated. At the $N - 1$ loci stage, a $2^{N-2} \times 2^{N-2}$ matrix $\mathbf{A}_{N-1}$ is defined by

$$\mathbf{A}_{N-1} = \left((-1)^{\boldsymbol{\xi}_i^T \boldsymbol{\xi}_j}\right), \tag{2.6.7}$$

where the rows and columns of $\mathbf{A}_{N-1}$ are ordered according to the ordering of the vectors in the set of gametic symbols. Let $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$ denote a $2^{N-2} \times 1$ column vectors of λ parameters and gametic probabilities, respectively. Then, given the matrix $\mathbf{A}_{N-1}$ for the case of $N-1$ loci, the numerical values of the λ parameters would be determined and defined by the vector-matrix equation

$$\boldsymbol{\lambda} = 2\mathbf{A}_{N-1}\boldsymbol{\gamma}. \tag{2.6.8}$$

To proceed recursively from the case of $N-1$ loci to the case of N loci, derive a set of 2^{N-1} gametic symbols by placing a 0 in position N for each of the symbols in the set $\mathbb{G}_{N-1}$ Then perform a similar operation by placing a 1 in position $N-1$ and a 0 in position N for each of the symbols in $\mathbb{G}_{N-1}$. The set $\mathbb{G}_N$ is then the union of the two sets just described and contains 2^{N-1} gametic symbols, which can also be interpreted as N dimensional vectors of zeros and ones. The matrix $\mathbf{A}_N$ and the vector of λ parameters would be defined as above for the case of $N-1$ loci.

For the general case of N linked loci, the complementary symbols of those in the set $\mathbb{G}_N$ may be expressed in a succinct form. Let $\mathbf{1}$ denote a gametic symbol containing all $1's$, indicating that a gamete of this type contains only paternal genes. Similarly, the symbol $\mathbf{0}$ denotes a gamete that contains all maternal genes. For any symbol $\xi \in \mathbb{G}_N$, its complementary symbol ξ^c is defined by the equation

$$\xi + \xi^c = \mathbf{1} \tag{2.6.9}$$

so that

$$\xi^c = \mathbf{1} - \xi. \tag{2.6.10}$$

Observe that $\mathbf{0}^c = \mathbf{1}$. Given this notation, the symmetry conditions of the gametic distribution may be expressed succinctly in a form by stating that for every $\xi \in \mathbb{G}_N$

$$\gamma(\xi) = \gamma(\xi^c). \tag{2.6.11}$$

We are now ready to state a theorem that provides a generalization of the cases of two and three linked loci to an arbitrary number of loci $N \geq 2$.

Theorem 2.6.1: (1) The sequence of matrices $\mathbf{A}_N$, $N = 2, 3, \ldots$ just defined satisfies the recursive equation

$$\mathbf{A}_N = \begin{bmatrix} \mathbf{A}_{N-1} & \mathbf{A}_{N-1} \\ \mathbf{A}_{N-1} & -\mathbf{A}_{N-1} \end{bmatrix} \tag{2.6.12}$$

for $N \geq 3$

(2) For every integer $N \geq 2$

$$\mathbf{A}_N^T = \mathbf{A}_N \tag{2.6.13}$$

so that $(\mathbf{A}_N)$, $N = 2, 3, \ldots$ is a sequence of symmetric matrices,

$$\mathbf{A}_N^2 = 2^{N-1}\mathbf{I}_{2^{N-1}}, \tag{2.6.14}$$

where $\mathbf{I}_{2^{N-1}}$ is an identity matrix of order 2^{N-1}, and the equation connecting the vectors $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ is

$$\boldsymbol{\gamma} = \frac{1}{2^N}\mathbf{A}_N\boldsymbol{\lambda}. \tag{2.6.15}$$

The proof of this theorem is given in the appendix.

With each lambda parameter $\lambda(\boldsymbol{\xi}_i)$ such that $i \neq 1$ there is associated a recombination probability defined by

$$\lambda(\boldsymbol{\xi}_i) = 1 - 2\rho((\boldsymbol{\xi}_i). \tag{2.6.16}$$

Equivalently,

$$\rho((\boldsymbol{\xi}_i) = \frac{1 - \lambda(\boldsymbol{\xi}_i)}{2}. \tag{2.6.17}$$

Given numerical values of the parameters $\lambda(\boldsymbol{\xi}_i)$, this formula could be used to calculate the recombination probabilities.

There is also another interesting representation of any parameter $\lambda(\boldsymbol{\xi}_i)$ such that $i \neq 1$. For the row corresponding to $\boldsymbol{\xi}_i$ in the vector $\mathbf{A}_N\boldsymbol{\gamma}$, let $S_1(\xi_i)$ denote the sum of the terms with positive signs and let $S_2(\xi_i)$ be the sum of the terms with negative signs. Then,

$$\lambda(\boldsymbol{\xi}_i) = 2\left(S_1(\xi_i) - S_2(\xi_i)\right) \tag{2.6.18}$$

for all $i \neq 1$. It is also of interest to note that

$$2\left(S_1(\xi_i) + S_2(\xi_i)\right) = 1, \tag{2.6.19}$$

because the sum of the elements of the gametic vector $2\boldsymbol{\gamma}$ is 1. From this result it also follows that $-1 \leq \lambda(\boldsymbol{\xi}_i) \leq 1$ for all $i \neq 1$. It is thus possible to derive another formula for the recombination probabilities by observing that

$$\begin{aligned} \rho((\boldsymbol{\xi}_i) &= \frac{1 - \lambda(\boldsymbol{\xi}_i)}{2} \\ &= \frac{2\left(S_1(\xi_i) + S_2(\xi_i)\right) - 2\left(S_1(\xi_i) - S_2(\xi_i)\right)}{2} \\ &= 2S_2(\xi_i)\,. \end{aligned} \tag{2.6.20}$$

The last equation expresses the recombination probability $\rho((\boldsymbol{\xi}_i)$ as a sum of gametic probabilities, but from the computational point of view, it is simpler to compute $\rho((\boldsymbol{\xi}_i)$ from $\lambda(\boldsymbol{\xi}_i)$.

2.7 Theoretical Calculations in Statistical and Population Genetics

In statistical and population genetics, it is often of interest to calculate selected probabilities from the linkage or gametic distribution when it is known that a set of genes is in the same linkage group. Consider, for example, a mating of the form

$$\frac{000}{111} \otimes \frac{000}{111}. \tag{2.7.1}$$

Then, it may be of interest to calculate the probability that an offspring from such a mating is homozygous with respect to some pattern ξ of recombination or non-recombination. As in the foregoing sections, when displaying a mating symbolically, it will be assumed that the genotype of the female in on the left and that of the male the right.

The same question could be asked if some larger number $N > 3$ of linked loci were under consideration. Let H denote the event that an offspring for such a mating is homozygous with respect to some pattern ξ. In the general case, given any pattern ξ, the probability that an offspring of such a mating is homozygous with respect to pattern $\xi \in \mathbb{G}_N$ is

$$\gamma^2(\xi). \tag{2.7.2}$$

Therefore, the probability an offspring from such a mating is homozygous with respect to patterns at N linked loci is

$$P[H] = \sum_{\xi} \gamma^2(\xi). \tag{2.7.3}$$

However, $\gamma(\xi) = \gamma(\xi^c)$ for all ξ so that

$$\begin{aligned} P[H] &= \sum_{\xi} \gamma^2(\xi) = 2 \sum_{\xi \in \mathbb{G}_N} \gamma^2(\xi) \\ &= 2\boldsymbol{\gamma}^T \boldsymbol{\gamma} \\ &= \frac{2}{(2^N)^2} \boldsymbol{\lambda}^T \mathbf{A}_N^T \mathbf{A}_N \boldsymbol{\lambda} \\ &= \frac{2^N}{(2^N)^2} \boldsymbol{\lambda}^T \boldsymbol{\lambda} = \frac{1}{2^N} \boldsymbol{\lambda}^T \boldsymbol{\lambda} \\ &= \frac{1}{2^N} \left(1 + \sum_{i \neq 1} \lambda^2(\xi_i) \right). \end{aligned} \tag{2.7.4}$$

From this equation, it can be seen that given a numerical specification of the vector $\boldsymbol{\lambda}$, the probability in question could be calculated.

By way of a numerical example, in the table below, numerical values for the recombination probabilities for the case of three linked loci have been assigned the values in column 2 for each pattern ξ. The corresponding λ-values are listed in column three of the table.

ξ	$\rho(\xi)$	$\lambda(\xi)$
000	0	1
100	0.25	0.50
010	0.35	0.30
110	0.10	0.80

(2.7.5)

For the values of the recombination probabilities in this table, the probability an offspring from a mating of the type under consideration is homozygous with respect to some pattern ξ has the value

$$P\left[H\right] = 0.2475. \tag{2.7.6}$$

On the other hand, if the three loci assorted independently, then this probability would have the value

$$P\left[H\right] = \frac{1}{8} = 0.125. \tag{2.7.7}$$

As one would expect, as a general rule, when a set of loci is in the same linkage group, then the probability that an individual from a mating of the type under consideration is homozygous with respect to some pattern ξ is greater than if the loci assorted independently.

In some simple cases, the event H may be interpreted as homozygous with respect to a set of alleles at three or more loci. For example, at the Mendelian level, let A, B C be dominant alleles at three linked loci, and let a, b and c denote the corresponding recessive alleles. Then, if in the above example the correspondence

$$\frac{000}{111} \leftrightarrow \frac{ABC}{abc} \tag{2.7.8}$$

holds, then, as before, the probability that an offspring of this mating in homozygous with respect to the two alleles at the three loci under consideration is

$$P\left[H\right] = 0.2475. \tag{2.7.9}$$

The illustrative calculations outlined above are just a few of the many calculations that could be carried out using specified numerical version of the linkage or gametic distribution. For example, suppose for some set of $N \geq 2$ linked loci, the set of recombination probabilities

$$(\rho(\xi) \mid \xi \in \mathbb{G}_N) \tag{2.7.10}$$

has been estimated or specified numerically, and suppose the patterns are ordered in the manner described in the foregoing section. Then, it would be straight forward to write software to compute the set

$$(\lambda(\xi) = 1 - 2\rho(\xi) \mid \xi \in \mathbb{G}_N) \tag{2.7.11}$$

of corresponding λ-values and shape them into a column vector $\boldsymbol{\lambda}$ in this prescribed order. Furthermore, by using the recursive equation for computing a matrix $\mathbf{A}_N$ for the case of N linked loci, it would be possible to compute a numerical version of the gametic distribution, using the formula

$$\boldsymbol{\gamma} = \frac{1}{2^N}\mathbf{A}_N\boldsymbol{\lambda}. \tag{2.7.12}$$

The next step in the process of developing a procedure to calculate events of interest is to find a way of computing the probability of finding an individual with genotype (ξ_f, ξ_m) is an offspring of mating of the type under consideration for the case of $N \geq 2$ linked loci. As a first step in this process, let $\mathbb{G}^c$ denote a set of patterns such that $\xi^c \in \mathbb{G}^c$ if, and only if, $\xi \in \mathbb{G}$. Then, because it is assumed that meiosis is a balanced process, for every $\xi^c \in \mathbb{G}^c$ let $\gamma(\xi^c) = \gamma(\xi)$ for every $\xi \in \mathbb{G}_N$ and let $\boldsymbol{\gamma}_c$ denote the $2^{N-1} \times 1$ vector

$$\boldsymbol{\gamma}_c = (\gamma(\xi^c) \mid \xi^c \in \mathbb{G}^c). \tag{2.7.13}$$

Observe that this vector is merely a relabeling of the elements of $\boldsymbol{\gamma}$ so that no additional computations are required. To find a matrix containing the desired probabilities, it will be necessary to introduce the $2^N \times 1$ partitioned vector $\boldsymbol{g}$ defined by

$$\boldsymbol{g} = \begin{pmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\gamma}_c \end{pmatrix} = (g(\xi) \mid \xi \in \mathbb{G} \cup \mathbb{G}^c). \tag{2.7.14}$$

Then, the $2^N \times 2^N$ matrix defined by

$$\boldsymbol{\Gamma} = \left(P\left[(\xi_f, \xi_m)\right] = g(\xi_f)\, g(\xi_m) \mid \xi_f \in \mathbb{G} \cup \mathbb{G}^c,\ \xi_m \in \mathbb{G} \cup \mathbb{G}^c\right) \tag{2.7.15}$$

contains all the probabilities of interest.

The matrix $\boldsymbol{\Gamma}$ also may also be represented as the partitioned matrix

$$\boldsymbol{\Gamma} = \boldsymbol{g}\boldsymbol{g}^T = \begin{bmatrix} \boldsymbol{\gamma}\boldsymbol{\gamma}^T & \boldsymbol{\gamma}\boldsymbol{\gamma}_c^T \\ \boldsymbol{\gamma}_c\boldsymbol{\gamma}^T & \boldsymbol{\gamma}_c\boldsymbol{\gamma}_c^Y \end{bmatrix}. \tag{2.7.16}$$

Observe that the $2^{N-1} \times 2^{N-1}$ in the upper left partition of $\mathbf{\Gamma}$ has the form

$$\boldsymbol{\gamma}\boldsymbol{\gamma}^T = \left(\gamma\left(\xi_f\right)\gamma\left(\xi_m\right) \mid \xi_f \in \mathbb{G}_N,\ \xi_m \in \mathbb{G}_N\right); \tag{2.7.17}$$

whereas that in the upper right partition has the form

$$\boldsymbol{\gamma}\boldsymbol{\gamma}_c^T = \left(\gamma\left(\xi_f\right)\gamma_c\left(\xi_m\right) \mid \xi_f \in \mathbb{G}_N,\ \xi_m \in \mathbb{G}_N^c\right). \tag{2.7.18}$$

The other sub-matrices in this partitioned matrix have similar interpretations. To expedite the computation of such matrices of probabilities, it would be helpful to use some software package that has pre-programmed operations on matrices included in the underlying code. Three such packages are MATLAB, APL and S-PLUS . Such packages also contain commands to select desired elements from the matrix $\mathbf{\Gamma}$ going into the calculation of the probabilities that offspring from matings of the type under consideration will have properties of interest

An example of such a matrix operation is that of computing the trace of a square matrix $\mathbf{\Gamma}$, which is the probability that an offspring in homozygous respect to the patterns of recombination under consideration. Let $\mathbf{A} = (a_{ij})$ be any $n \times n$ matrix. By definition, the trace of the matix $\mathbf{A}$ is

$$tr\mathbf{A} = \sum_{i=1}^{n} a_{ii}. \tag{2.7.19}$$

From this observation, it follows that the probability that an offspring from a mating of the type under consideration is homozygous with respect to some pattern ξ is

$$P\left[H\right] = tr\,\mathbf{\Gamma} = \ tr\boldsymbol{\gamma}\boldsymbol{\gamma}^T + tr\boldsymbol{\gamma}_c\boldsymbol{\gamma}_c^Y = 2tr\boldsymbol{\gamma}\boldsymbol{\gamma}^T = 2\boldsymbol{\gamma}^T\boldsymbol{\gamma}. \tag{2.7.20}$$

There are at least two other situations that arise in statistical and population genetics in which numerical versions of the gametic or linkage distribution for two or more loci could be applied. When analyzing pedigree data in statistical genetics, for example, it is often of interest to calculate the probability that an individual is homozygous by descent in the sense that at some set of linked loci an individual is homozygous by virtue of possessing two copies at each locus of genes inherited from an ancestor in his pedigree. Calculating such probabilities can be complicated and, therefore, no examples of these calculations with be undertaken here. In evolutionary population genetics, it would, for example, be of interest to study rates of convergence to linkage equilibrium for the case of many loci under random mating with no selection or mutation. As will be illustrated in the next chapter, numerical versions of the linkage distribution for several loci would be very useful in such studies.

There is another complication that may arise in the illustrative calculations described in this section; namely, the case in which recombination probabilities in females and males are different. Fortunately, such a complication may easily be incorporated into the calculations described above. For example, let $\boldsymbol{\gamma}_f$ and $\boldsymbol{\gamma}_m$ denote, respectively, the female and male linkage distributions. Both of these vectors could be calculated as described above, given a set of recombination probabilities for each sex. Given the vectors $\boldsymbol{\gamma}_f$ and $\boldsymbol{\gamma}_m$ the calculations could be described above could be done separately for females and males to compute two vectors $\boldsymbol{g}_f$ and $\boldsymbol{g}_m$. The matrix $\boldsymbol{\Gamma}$ of the desired probabilities would then have the form

$$\boldsymbol{\Gamma} = \boldsymbol{g}_f \boldsymbol{g}_m^T. \tag{2.7.21}$$

2.8 Appendix: Proof of Theorem 2.6.1

Proof: (1) For the sake of convenience four cases, corresponding to the sub-matrices,

$$\mathbf{A}_N = \begin{bmatrix} \mathbf{A}_{N-1} & \mathbf{A}_{N-1} \\ \mathbf{A}_{N-1} & -\mathbf{A}_{N-1} \end{bmatrix} \tag{2.8.1}$$

will be considered. Case 1 is the sub-matrix in the upper left. In case 1, a zero has been added in the *N-th* position of both vectors $\boldsymbol{\xi}_i$ and $\boldsymbol{\xi}_j$. Hence, the number

$$(-1)^{\boldsymbol{\xi}_i^T \boldsymbol{\xi}_j} \tag{2.8.2}$$

does not change sign as one proceeds form the case of $N-1$ loci to the case of N loci. The sub-matrix in case 1 is, therefore, $\mathbf{A}_{N-1}$.

Case 2 is the sub-matrix on the upper right. In this case, a zero has been inserted in positions $N-1$ and N of the vector $\boldsymbol{\xi}_i$ and a one has been inserted into position $N-1$ and a zero in position N of the vector $\boldsymbol{\xi}_j$. Therefore, the sign of the number does not change as one proceeds from the case of $N-1$ loci to the case of N loci so that the sub-matrix in this case is $\mathbf{A}_{N-1}$. Case 3 is the sub-matrix in the lower left. That this sub-matrix is $\mathbf{A}_{N-1}$ follows by symmetry from case 2.

Case 4 in the sub-matrix on the lower right. In this case, a one has been inserted in position $N-1$ and a zero in position N for both vectors $\boldsymbol{\xi}_i$ and $\boldsymbol{\xi}_j$. Therefore, in this case the number in changes sign so that the sub-matrix in the lower right becomes $-\mathbf{A}_{N-1}$ as one proceeds form the case of $N-1$ loci to the case of N loci. This completes the proof of assertion (1).

Assertion (2) of the theorem is proven by induction on N. For $N = 2$ we see that

$$\mathbf{A}_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \tag{2.8.3}$$

is a symmetric matrix. By the induction hypothesis, assume $\mathbf{A}_{N-1}$ is a symmetric matrix. Then, from the recursive equation, it follows that

$$\begin{aligned} \mathbf{A}_N^T &= \begin{bmatrix} \mathbf{A}_{N-1}^T & \mathbf{A}_{N-1}^T \\ \mathbf{A}_{N-1}^T & -\mathbf{A}_{N-1}^T \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{N-1} & \mathbf{A}_{N-1} \\ \mathbf{A}_{N-1} & -\mathbf{A}_{N-1} \end{bmatrix}. \\ &= \mathbf{A}_N \end{aligned} \tag{2.8.4}$$

This proves equation is valid for all $N \geq 2$.

To prove that

$$\mathbf{A}_N^2 = 2^{N-1}\mathbf{I}_{2^{N-1}} \tag{2.8.5}$$

for all $N \geq 2$, observe that

$$\mathbf{A}_2^2 = 2\mathbf{I}_2, \tag{2.8.6}$$

and, by the induction hypothesis suppose

$$\mathbf{A}_{N-1}^2 = 2^{N-2}\mathbf{I}_{2^{N-2}}. \tag{2.8.7}$$

Then,

$$\begin{aligned} \mathbf{A}_N^2 &= \begin{bmatrix} \mathbf{A}_{N-1} & \mathbf{A}_{N-1} \\ \mathbf{A}_{N-1} & -\mathbf{A}_{N-1} \end{bmatrix} \begin{bmatrix} \mathbf{A}_{N-1} & \mathbf{A}_{N-1} \\ \mathbf{A}_{N-1} & -\mathbf{A}_{N-1} \end{bmatrix} \\ &= \begin{bmatrix} 2\mathbf{A}_{N-1}^2 & \mathbf{0} \\ \mathbf{0} & 2\mathbf{A}_{N-1}^2 \end{bmatrix} = 2^{N-1}\mathbf{I}_{2^{N-1}}. \end{aligned} \tag{2.8.8}$$

This proves that the equation is valid for all $N \geq 2$.

Finally, the equation

$$\boldsymbol{\gamma} = \frac{1}{2^N}\mathbf{A}_N\boldsymbol{\lambda} \tag{2.8.9}$$

follows by solving the equation

$$\boldsymbol{\lambda} = 2\mathbf{A}_N\boldsymbol{\gamma} \tag{2.8.10}$$

for the vector $\boldsymbol{\gamma}$, using the orthogonal properties of the matrix $\mathbf{A}_N$. Also observe that if $\lambda(\xi_i) = 0$ for all $i \neq 1$, then

$$\boldsymbol{\gamma} = \frac{1}{2^N}\mathbf{1}_{N-1}, \tag{2.8.11}$$

where $\mathbf{1}_{N-1}$ is a $(N-1)\times 1$ vector of ones, which is the case of independent assortment for N loci.

Bibliography

[1] Bernoulli, J. (1713) Ars Conjectandi, pars quarta.

[2] Bronowski, J. (1974) **The Ascent of Man**. Little Brown and Company, Boston, Toronto.

[3] Browning, S. (2000) The Relationship Between Count - Location and Stationary Renewal Models for the Chiasma Process. Genetics **155**:1955–1960.

[4] Church, G. M. (2006) Genomes For All. Scientific American **294**:46–54.

[5] Coop, G. and Przeworski, M. (2007) An Evolutionary View of Human Recombination. Nature Reviews/Genetics **8**:23–34.

[6] Coop, G., Wen, X., Ober, C., Pritchard, J. K. and Przeworski, M. (2008) High-Resolution Mapping of Crossovers Reveals Extensive Variation in Fine-Scale Recombination Patterns Among Humans. Science **319**:1395–1398.

[7] Devlin, K. (2000) **The Math Gene**. Basic Books, Perseus Books Group.

[8] Ewens, W. J. and Grant, G. R. (2005) **Statistical Methods in Bioinformatics**, Second Edition, Springer.

[9] Liu, B. H. (1998) **Statistical Genomics, Linkage, Mapping and QTL Analysis**. CRC Press, Boca Raton and New York.

[10] Raymond, C. K. *et al.* (2005) Ancient haplotypes of the HLA Class II region. Genome Research **15**:1250–1257.

[11] Schnell, F. W. (1961) Some General Formulations of Linkage Effects in Inbreeding. Genetics **46**:947–957.

[12] Sinnott, E. W., Dunn, L. C. and Dobzhansky, Th. (1950) **Principles of Genetics**, McGraw Hill, New York, Toronto and London.

[13] Strachan, T. and Read, A. P. (2004) **Human Molecular Genetics**, Third Edition. Garland Science, Taylor and Francis Group, London and New York.

[14] Thompson, E. A. (2000) **Statistical Inference From Genetic Data on Pedigrees**. Institute of Mathematical Statistics, Beachwood, Ohio.

[15] Uspensky, J. V. (1937) **Introduction to Mathematical Probability**. McGraw Hill, New York and London.

[16] Venter, J. C. (2007) **A Life Decoded- My Genome: My Life**. Penguin Books, London.

[17] Zhang, K., Zhu, J., Shendure, J., Porreca, G. J., Aach, J. D., Mitra, R. D. and Church, G. M. (2006) Long - Range Polony Haplotyping of Individual Chromosome Molecules. Nature Genetics **38**:382–387.

[18] Zhao, H. and Speed, T. P. (1996) On Genetic Map Functions. Genetics **142**:1369–1377.

Chapter 3

Linkage and Recombination in Large Random Mating Diploid Populations

3.1 Introduction

The purpose of this chapter is to study some aspects of the dynamics of biological populations in terms of their genetic structure with respect to multiple linked loci, when mating is random but there is no selection or mutation. Furthermore, attention will be restricted to large diploid populations with two sexes, females and males. At the outset it should be stated that the theory which follows has at least one serious defect; namely it does not take into account population number. It is widely recognized that population number plays an important role in the dynamics of a biological population, particularly when population size is small or when the number of possible genotypes exceeds that of the human population of the world. Consequently, any mathematical model describing the evolution of a biological population should take into account the number of individuals present in the population at any time.

It is perhaps unfortunate that the theory of the present chapter is restricted to populations with non-overlapping generations as exemplified by annual plants. Consequently, the theory in its present form is not applicable, in a strict sense, to human populations, for in human populations generations are overlapping, i.e., all ages are represented in the population simultaneously. Despite these defects and restrictions, however, the theory which follows is interesting, seems to have some validity in large populations, and will very likely serve as a stepping stone to the case of populations with over lapping generations. Even though these limitations are widely recognized, the theories and conceptual metaphors, presented in this chapter form a basis for a great deal of mathematical genetics as it exists today and as it is applied in quantitative and human genetics as well as in searches of the human genome for signatures of natural selection, which are thought to be useful in searching for genes implicated in the expression of observed phenotypes such as resistance to disease. In this connection, the recent paper of Hinds *et al.* (2005) where the detection of linkage disequilibrium in a genome wide search was interpreted as a possible signature of natural selection.

Briefly, the main objective of this chapter is to show that in large random mating populations in which there is no mutation or selection with respect to a set of linked loci, convergence to a linkage equilibrium occurs as the number of generation becomes large. Various books on population genetics contain results on the convergence to a linkage equilibrium with respect to two or three linked loci, see, for example, Crow and Kimura (1970), Cavalli-Sforza and Bodmer (1971), Buerger (2000), Christiansen (2000) and Ewens (2004). However, none of these books contain an account of convergence to a linkage equilibrium as some number of linked loci $N \geq 4$ at the level of generality presented in this chapter. As a first step towards the development of general theories, attention will be focused on a population with respect to one autosomal locus with an arbitrary but finite number of alleles. As this chapter is devoted to a review of results from classical population genetics, which for the most part have been known for several decades, the presentation will in terms of more formal mathematics than in the other chapters of this book.

3.2 The One Locus Case

Arbitrary alleles at the locus under consideration will be represented by lower case letters toward the end of the alphabet, e.g., u, v, w, x, y, and z. An arbitrary genotype will be represented by an ordered pair (x, y), where the first and second members of the pair is the allele received from maternal and paternal parent respectively. For fixed alleles x and y the genotypes (x, y) and (y, x) are identical from the biological point of view, but for mathematical purposes it will be convenient to distinguish between them. If there are s alleles at the locus under consideration, then there are s homozygotes of the form (x, x), $s(s - 1)$ ordered pairs of the form (x, y) $(x \neq y)$, but only $s + s(s - l)/2 = s(s + l)/2$ biologically distinguishable genotypes.

In order to fix ideas, consider a large population consisting of two sexes. In what follows, the sex of an individual will not be specified, but later on we shall show that the results may be easily modified to take into account the presence of two sexes in the population. If a population is very large, then it is reasonable to speak of the probability of the event a member of the population is a representative of genotype (x, y). Let $P^{(n)}(x, y)$ be the probability a. Member of the population is a representative of genotype (x, y) generation $n = 0, 1, 2, \ldots$. Because the genotypes (x, y) and (y, x) are indistinguishable biologically, it will be required that

$$P^{(n)}(x, y) = P^{(n)}(y, x) \tag{3.2.1}$$

for all generations $n = 0, 1, 2, \ldots$

Thus if $x \neq y$, then the probability a member of the population carries alleles x and y is $P^{(n)}(x, y)$ $+P^{(n)}(y, x)$; while if $x = y$, then the probability a member of the population carries two copies of the allele x is $P^{(n)}(x, x)$. The set of probabilities

$$\left(P^{(n)}(x, y) \mid (x, y)\right), \tag{3.2.2}$$

where x and y range over all alleles present at the locus under consideration, will henceforth be called the genotypic distribution in generation n. It will be noted that the genotypic distribution satisfies the conditions

$$P^{(n)}(x, y) \geq 0$$

for all $n \geq 0$, x, y, and

$$\sum_x \sum_y P^{(n)}(x, y) = 1, \tag{3.2.3}$$

where the sum extends over all alleles present at the locus under consideration.

Let us next direct our attention to the set of all possible matings in the population with respect to the locus under consideration. In each mating, a female of genotype say (u, v) will be mated to a male of genotype say (x, y), and each such mating will occur with a certain probability. Let $P^{(n)}((u, v); (x, y))$ be the probability a female of genotype(u, v) is mated to a male of genotype (x, y) in generation n. The set of all probabilities

$$\left(P^{(n)}((u, v); (x, y)) \mid u, v, x, y\right),$$

where u, v, x, and y range over all alleles present at the locus, is called the mating distribution in generation n. This distribution satisfies conditions comparable to condition (3.2.3). An important special case arises when the mating distribution is determined by the genotypic distribution in a simple way.

Definition 3.2.1: The population is said to mate at random in generation n if, and only if,

$$P^{(n)}((u,v);(x,y)) = P^{(n)}\left(u,v\right)P^{(n)}\left(x,y\right)$$

for all choices of alleles u, v, x, and y at the locus under consideration. In other words, in a random mating population the mating distribution is the product of the genotypic distributions. As we shall see in the sequel this definition has rather far reaching consequences.

The following simple example is at the root of the results of this chapter. Consider a population with respect to one locus with two alleles u and v and let $P^{(0)}(u,u) = D$, $2P^{(0)}(u,v) = 2H$, and $P^{(0)}(v,v) = R$ be the genotypic distribution in the initial generation. A question that naturally arises is: if there is no selection or mutation and the population mates at random, then what is the genotypic distribution in generation one? The meaning of the term "no selection" will be made clear later and by no mutation we mean neither allele u nor v changes into the other when parents transmit these alleles to their offspring. The situation we have in mind may be conveniently described by the table which appears below.

$M =$ Mating	$P\left(M\right)$	$P\left(u,u\right)$	$P\left(u,v\right)$	$P\left(v,v\right)$
$(u,u)\otimes(u,u)$	D^2	1	0	0
$(u,u)\otimes(u,v)$	$4DH$	$\frac{1}{2}$	$\frac{1}{2}$	0
$(u,u)\otimes(v,v)$	$2DR$	0	1	0
$(u,v)\otimes(v,v)$	$4HR$	0	$\frac{1}{2}$	$\frac{1}{2}$
$(v,v)\otimes(v,v)$	R^2	0	0	1
$(u,v)\otimes(u,v))$	$4H^2$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

(3.2.4)

In table (3.2.4), the genotypes of the male and female in a mating are not distinguished. For example, the mating $(u,u)\otimes(u,v)$ includes females of genotype (u,u) mated to males of genotype (u,v) as well as females of genotype (u,v) mated to males of genotype (u,u). The assumption that no mutations occur is expressed by the zeros which appear in the last three columns of the table. By no selection we mean the genotypic distribution in generation one is obtained by multiplying the second column by the third, fourth, and fifth columns and adding, which leads to the following equations.

$$\begin{aligned} P^{(1)}\left(u,u\right) &= D^2 + 2DH + H^2 = \left(D+H\right)^2 \\ 2P^{(1)}\left(u,v\right) &= 2\left(DH + DR + HR + H^2\right) = 2\left(D+H\right)\left(H+R\right) \quad (3.2.5) \\ P^{(1)}\left(v,v\right) &= 2HR + R^2 + H^2 = \left(H+R\right)^2 \end{aligned}$$

It will be noted that the quantities

$$\begin{aligned} p(u) &= P^{(0)}(u,u) + P^{(0)}(u,v) = D + H \\ p(v) &= P^{(0)}(u,v) + P^{(0)}(v,v) = H + R \end{aligned} \tag{3.2.6}$$

occur in equations (3.2.6). The quantities $p(u)$ and $p(v)$ may be interpreted as the probability of finding alleles u and v, respectively, in the gene pool of the population in the initial generation. These observations lead to a result which is known nowadays as the Hardy-Weinberg Law.

Theorem 3.2.1: If the population mates at random and there is no selection or mutation in generations $n = 0, 1, 2, \ldots, N$, where N is an arbitrary positive integer, then the following equations

$$\begin{aligned} P^{(n)}(u,u) &= p^2(u) \\ 2P^{(n)}(u,v) &= 2p(u)\,p(v) \\ P^{(n)}(v,v) &= p^2(v) \end{aligned} \tag{3.2.7}$$

for $n = 1, 2, \ldots, N$. Moreover, the equation

$$\left(P^{(n)}(u,v)\right)^2 = P^{(n)}(u,u)\,P^{(n)}(v,v) \tag{3.2.8}$$

holds for $n = 1, 2, \ldots, N$.

Proof: For $n = 1$ equations (3.2.7) reduce to equations (3.2.5) and (3.2.6). The truth of equations (3.2.7) for arbitrary n may be shown by mathematical induction. Suppose equations (3.2.7) are true for some integer n such that $1 \le n \le N$. Then we may carry out the computations of table 3.2.4 with $D = p^2(u)$, $H = p(u)p(v)$ and $R = p^2(v)$. Proceeding as above, it follows that

$$\begin{aligned} P^{(n+1)}(u,u) &= D^2 + 2DH + R^2 = (D+H)^2 = p^2(u) \\ 2P^{(n+1)}(u,v) &= 2\left(DH + DR + HR + H^2\right) \\ &= 2(D+H)(H+R) \\ &= 2p(u)\,p(v) \\ P^{(n+1)}(v,v) &= 2HR + R^2 + H^2 = (H+R)^2 = p^2(v)\,, \end{aligned} \tag{3.2.9}$$

which proves (3.2.7). Assertion (3.2.8) is an easy consequence of equations (3.2.7), and this completes the proof of the theorem.

It should be noted that an equilibrium in the genotypic distribution is reached in the first generation under random mating and no selection or mutation, and, furthermore, this equilibrium is maintained as long as the hypotheses of the theorem hold. The positive integer N is inserted into the theorem to emphasize the feeling that it seems reasonable to assume that the hypotheses of the theorem would be expected to hold only for some finite number of generations. From the mathematical point of view the positive integer N is unnecessary in Theorem 3.2.1. As we shall now see, Theorem 3.2.1 easily carries over to the case of an arbitrary number of alleles at a locus.

Let $p^{(n)}(x)$ be the probability of finding allele x in the gene pool of the population in generation n. The set of all probabilities $\left(p^{(n)}(x) \mid x\right)$, where x ranges over all alleles present at the locus, is called the gametic distribution in generation n. As of symmetry condition (3.2.1) and the assumption that no mutations occur, it follows that

$$p^{(n)}(x) = \sum_{y} P^{(n)}(x, y), \tag{3.2.10}$$

where the sum extends over all alleles present at the locus under consideration. In the absence of mutation., an offspring of genotype (u, v) may be produced by any of the four matings $(u, z) \otimes (v, w)$, $(z, u) \otimes (v, w)$, $(u, z) \otimes (w, v)$, and $(z, u) \otimes (w, v)$, where w and z are arbitrary alleles. Each mating produces an offspring of genotype (u, v) with probability 1/4. From symmetry condition (3.2.1), it follows that under random mating with no selection or mutation, the probability of finding an individual of genotype (u, v) in generation $n + 1$ is

$$\begin{aligned} P^{(n+1)}(u, v) &= \sum_{z} \sum_{w} P^{(n)}(u, z) P^{(n)}(v, w) \\ &= \left(\sum_{z} P^{(n)}(u, z)\right) \left(\sum_{w} P^{(n)}(v, w)\right) \qquad (3.2.11) \\ &= p^{(n)}(u) p^{(n)}(v). \end{aligned}$$

Therefore, for $n = 0$ it follows that

$$P^{(1)}(u, v) = p^{(0)}(u) p^{(0)}(v).$$

By using mathematical induction, it can be shown that

$$P^{(n)}(u, v) = p^{(0)}(u) p^{(0)}(v) \tag{3.2.12}$$

for every $n = 1, 2, \ldots, N$ and arbitrary alleles u and v, which is the desired extension Theorem (3.2.1) to the case of an arbitrary number of alleles at a locus.

From the theory which has been developed so far, we see random mating in the absence of selection and mutation is equivalent to gametes of generation n uniting at random to produce the genotypes of generation $n+1$ in the sense that the genotypic distribution in generation $n+1$ is determined by the product of the gametic distribution in generation n. If it is assumed that mutations occur and there is no selection, then the genotypic distribution in generation $n+1$ is again determined by the product of the gametic distribution in generation n. Unlike the situation described in Theorem 3.2.1, however, an equilibrium in the genotypic distribution may be reached only in the limit.

To take mutation into account, with each genotype (u, v), we associate a gametic output distribution $q((u, v); x)$, where x extends over all alleles present at the locus under consideration and $q((u, v); x)$ is the conditional probability genotype (u, v) produces a gamete containing the allele x. It will be assumed the gametic output distribution is constant from generation to generation. The conditional probability $q((u, v); x)$ could also be expressed in terns of the probabilities alleles u and v mutate to allele x but we shall not go into these matters until mutation is discussed more fully.

Clearly, in the present set up the probability $p^{(n)}(x)$ of finding allele x in the gene pool of the population in generation n is given by

$$p^{(n)}(x) = \sum_{(u,v)} P^{(n)}\,(u, v)\, q((u, v); x) \tag{3.2.13}$$

where the sum extends over all possible genotypes at the locus under consideration. Under the hypothesis of random mating, the probability that in generation n a female of genotype (x, y) is mated to a male of genotype (z, w) and this mating produces an offspring of genotype (u, v) is

$$\begin{aligned} P^{(n+1)}\,(u, v) &= \sum_{(x,y)} \sum_{(z,w)} P^{(n)}\,(x, y)\, P^{(n)}\,(x, w)\, q((x, y); u) q((z.w); v) \\ &= \left(\sum_{(x,y)} P^{(n)}\,(x, y)\, q((x, y); u) \right) \left(\sum_{(z,w)} P^{(n)}\,(x, w)\, q((z.w); v) \right) \\ &= p^{(n)}\,(u)\, p^{(n)}\,(v) \end{aligned} \tag{3.2.14}$$

Therefore, when the mating is random with mutation but no selection, the genotypic distribution in generation $n+1$ is determined by the product of the gametic distribution in generation n. Observe in this derivation it has been tacitly assumed that the mutation process is independent in the probabilistic sense among individuals in a population. It is important to observe from (3.2.14) that the limiting behavior of the genotypic distribution is determined by the way in which the model of the gametic output distribution is formulated. If this distribution is formulated in terms of Markov chains, then the limiting behavior will easily from Markov chain theory. We shall, however, defer limit aspects of (3.2.14) until we have a portion of Markov chain theory at our disposal.

So far in our consideration of a population with two sexes, we have not specified the sex of an individual and have, in fact, assumed that the genotypic probabilities in the male and female populations were the same. We shall now show, under certain conditions to be stated later, that if the genotypic distributions in the male and female populations differ in the initial generation, then they are equalized after one generation of random mating with no mutation or selection.

Let $Q^{(n)}(x,y)$ and $P^{(n)}(x,y)$ be the probabilities of a female and male, respectively, are of genotype (x,y) in generation n, and let $q^{(n)}(x)$ and $p^{(n)}(x)$ be the probabilities of allele x in the gene pool of the female and male population, respectively. We shall again assume that mutations occur and that the gametic output distribution is the same in both sexes. Under this last assumption the probability $q^{(n)}(x)$ is defined by replacing $P^{(n)}(x,y)$ with $Q^{(n)}(x,y)$ in (3.2.13). It will also required that the conditions

$$\begin{aligned}\sum_{(x,y)} Q^{(n)}(x,y) &= 1\\ \sum_{(x,y)} P^{(n)}(x,y) &= 1\end{aligned} \tag{3.2.15}$$

hold for all $n = 0,1,2,\ldots$. In what follows, let p be the probability an offspring is female and let $q = 1-p$ be the probability it is male.

Given these conditions, if the mating is random., then the probability that in the initial generation a female of genotype (x,y) is mated with a male of genotype (z,w) and this mating produces a female offspring of genotype (u,v) is

$$pQ^{(0)}\left(x,y\right)q\left(\left(x,y\right);u\right)P^{(0)}(z,w)q\left(\left(z,w\right);v\right) \tag{3.2.16}$$

By summing over all genotypes (x,y) and (z,w), it follows that

$$Q^{(1)}\left(u,v\right) = pq^{(0)}\left(u\right)p^{(0)}\left(v\right).$$

Similarly,

$$P^{(1)}\left(u,v\right) = qq^{(0)}\left(u\right)p^{(0)}\left(v\right).$$

Therefore,

$$\frac{Q^{(1)}\left(u,v\right)}{p} = \frac{P^{(1)}\left(u,v\right)}{q} = q^{(0)}\left(u\right)p^{(0)}\left(v\right) = P_w^{(1)}\left(u,v\right) \tag{3.2.17}$$

so that within the female and male populations in the first generation the genotypic distribution is the same as denoted by the symbol $P_w^{(1)}(u,v)$. In this case, it is interesting to note that the symmetry condition

$$P_w^{(1)}(u,v) = P_w^{(1)}(v,u)$$

holds for all alleles u and v.

3.3 The Case of Many Autosomal Loci with Arbitrary Linkage

The purpose of this section is to generalize many of the results of the previous section to the case of an arbitrary number of loci with an arbitrary number of alleles at each locus. For the sake of concreteness let N be the number of loci under consideration and suppose there are r_k, $k = 1, 2, \ldots, N$, alleles at the *k-th* locus. An arbitrary genotype will be represented by the ordered pair of vectors $(\boldsymbol{x}, \boldsymbol{y})$. where $\boldsymbol{x} = (x_1, \ldots, x_N)$, $\boldsymbol{y} = (y_1, \ldots, y_N)$ and the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ represent the genes received from the maternal and paternal parent respectively. Either of the symbols x_k or y_k may represent any of the r_k alleles present at the *k-th* locus. Just as before the genotypes $(\boldsymbol{x}, \boldsymbol{y})$ and $(\boldsymbol{y}, \boldsymbol{x})$ are identical from the genetic point of view, but it will be convenient to distinguish between them mathematically.

In the sequel, the N dimensional vectors $\boldsymbol{x} = (x_1, \ldots, x_N), \boldsymbol{y} = (y_1, \ldots, y_N)$ and $\boldsymbol{z} = (z_1, \ldots, z_N)$ will represent any of the possible gametes generated by the population. As an example of the notation discussed in the previous paragraph suppose $N = 6$ and the alleles of each locus are numbered arbitrarily. The genotype $((1,2,3,4,5,6),(2,4,6,8,10,12))$ has alleles 1 and 2 at the first locus, 2 and 4 at the second locus, and so on down:n to 6 and 12 at the sixth locus. Moreover, he received the genes $(1,2,3,4,5,6)$ from his maternal parent and the genes $(2,4,6,8,10,12)$ from his paternal parent.

When one is considering linkage it is essential to consider the way in which the gametes enter the zygotes from which a mature individual originates. For example, the genotypes $((1,1),(2,2))$ and $((1,2),(2,1))$ will be identical in all respects except the probabilities with which they generate the gametes $(1,1),(1,2),(2,1)$, and $(2,2)$ will differ. This is, of course, an example of a coupling and repulsion situation mentioned in chapter two. More generally, if a population is considered with respect to N loci with r_k alleles at the *k-th* locus, then

$$s = \prod_{k=1}^{N} r_k \tag{3.3.1}$$

is the number of gametes that may be generated by the population. Accordingly, if genotypes $(\boldsymbol{x}, \boldsymbol{y})$ and $(\boldsymbol{y}, \boldsymbol{x})$ are considered identical, then when linkage is taken into account it will be necessary to distinguish $s(s+1)/2$ genotypes. Even if a moderate number of loci are considered, the number of genotypes which must be distinguished in the presence of linkage may be quite large. In fact, the number $s(s+1)/2$ could very well exceed to total size of a population. We shall, however, from now on assume that the size of a population far exceeds the number of distinguishable genotypes.

Before continuing with the general discussion, it will be helpful to consider an illustrative numerical example with respect to the *HLA-A*, *HLA-B* and *HLA-DRB*1 loci mentioned in chapter 2, which have 250, 500 and 300 known alleles, respectively. Therefore, with respect to these three loci, the number of gametes that may be generated in a population is

$$s = 250 \times 500 \times 300 = 37,\ 500,\ 000, \tag{3.3.2}$$

and the number of distinguishable genotypes in a population would be

$$(37500000)\,(37500000 + 1)\,/2 = 703,\ 125,\ 018,\ 750,\ 000. \tag{3.3.3}$$

It is interesting to observe that this number is of the order 10^{15} and far exceeds 6-8 $\times$ 10^9, which is approximately the range for size of the total human population on the earth as of 2006. This illustrative numerical example, not only helps make clear the need to accommodate the size of a population in genetic evolutionary models but also provides a glimpse of the large magnitude of potential genetic variation that may be present in human and other populations of diploid organisms with various systems of mating.

In the large hypothetical random mating population under consideration, let $P^{(n)}(\boldsymbol{x}, \boldsymbol{y})$ be the probability of genotype $(\boldsymbol{x}, \boldsymbol{y})$ in generation $n = 0, 1, 2, \ldots.$ Because the genotypes $(\boldsymbol{x}, \boldsymbol{y})$ and $(\boldsymbol{y}, \boldsymbol{x})$ are considered identical, it will be required that

$$P^{(n)}(\boldsymbol{x}, \boldsymbol{y}) = P^{(n)}(\boldsymbol{y}, \boldsymbol{x}) \tag{3.3.4}$$

for all nonnegative integers n and all choices of gametes x and y. It follows that the probability an individual in generation n carries the genes in the gametes $\boldsymbol{x}$ and $\boldsymbol{y}$ is $P^{(n)}(\boldsymbol{x}, \boldsymbol{y}) + P^{(n)}(\boldsymbol{y}, \boldsymbol{x})$. Similarly, the probability that an individual in generation n carries the genes in the gamete $\boldsymbol{x}$ in homozygous condition is $P^{(n)}(\mathrm{x}, \mathrm{x})$. The probability of gamete $\boldsymbol{x}$ in the gene pool of the population in generation n will be denoted by $p^{(n)}(\boldsymbol{x})$ and the collections of probabilities $(P^{(n)}(\boldsymbol{x}, \boldsymbol{y}))$ and $(p^{(n)}(\boldsymbol{x}))$, where $\boldsymbol{x}$ and $\boldsymbol{y}$ range over all possible gametes, will be referred to as the genotypic and gametic distributions, respectively, in generation n. Under the assumption of random mating and no selection there is a simple relation connecting the genotypic distribution of generation n with the gametic distribution in generation $n-1$. This relation is

$$P^{(n)}\left(\boldsymbol{x}, \boldsymbol{y}\right) = p^{(n-1)}\left(\boldsymbol{x}\right) p^{(n-1)}\left(\boldsymbol{y}\right) \tag{3.3.5}$$

which holds for all $n \geq 1$ and all choices of gametes $\boldsymbol{x}$ and $\boldsymbol{y}$, is an easy consequence of the discussion in section 3.2. This equation could also be taken as a definition of random mating.

In order to achieve our objectives, it will be necessary to consider the so-called marginal distributions associated with the gametic distribution. For example, the probability of a gamete carrying the allele x_1 at the first locus in generation n is

$$p_1^{(n)}\left(x_1\right) = \sum_{x_x} \cdots \sum_{x_N} p^{(n)}\left(x_1, x_2, \ldots, x_N\right), \tag{3.3.6}$$

and the probability of a gamete carrying alleles x_1 and x_2 in generation n is

$$p_{12}^{(n)}\left(x_1, x_2\right) = \sum_{x_3} \cdots \sum_{x_N} p^{(n)}\left(x_1, x_2, x_3, \ldots, x_N\right). \tag{3.3.7}$$

In general if we let S be the set of integers $(1, 2, \ldots, N)$, then with each subset A of S there corresponds a marginal probability $p_A^{(n)}\left(\boldsymbol{x}_A\right)$ defined as the sum

$$p_A^{(n)}\left(\boldsymbol{x}_A\right) = \sum_{[x_k \mid k \in A^c]} p^{(n)}\left(x_1, x_2, \ldots, x_N\right) \tag{3.3.8}$$

for all $n = 0, 1, 2, \ldots$. In this equation the symbol A^c stands for the complement of the subset A of S. The symbol $\boldsymbol{x}_A$ stands for a vector whose elements and dimensionality correspond to the elements in the set A. For example, if $A = (2, 4)$, then $\boldsymbol{x}_A = (x_2, x_4)$, where x_2 and x_4, may be any of the alleles at loci two and four. To ensure that $p_A^{(n)}\left(\boldsymbol{x}_A\right)$ is well defined for all subsets A of S, we let

$$p_A^{(n)}\left(\boldsymbol{x}_A\right) = p^{(n)}\left(\boldsymbol{x}\right) \tag{3.3.9}$$

if $A = S$, and let

$$p_A^{(n)}\left(\boldsymbol{x}_A\right) = 1 \tag{3.3.10}$$

if $A = \varphi$, the empty set. Observe that this assignment is consistent with the definition of $p_A^{(n)}(\boldsymbol{x}_A)$ because

$$p_\varphi^{(n)}(\boldsymbol{x}_\varphi) = \sum_{[x_k \mid k \in S]} p^{(n)}(x_1, x_2, \ldots, x_N) = 1. \tag{3.3.11}$$

As a final step in the preparation for stating a result that is fundamental when studying linkage and recombination in random mating populations, the linkage distribution studied in chapter 2 will be described with a slightly different notation. Let $\gamma(\xi)$ be any probability in the linkage distribution and suppose the pattern ξ is such that maternal genes occur at the subset A of S. Then set $\gamma(\xi) = \gamma(A)$, Observe that if the process of meiosis is balanced, then just as in chapter 2

$$\gamma(A) = \gamma(A^c). \tag{3.3.12}$$

These simple definitions and observations lead to a result of fundamental importance in the theory of random mating diploid populations with no selection or mutation.

Theorem 3.2.1: If the mating system is random and there is no mutation or selection, then for every generation $n \geq 1$

$$p^{(n)}(\boldsymbol{x}) = \sum_A \gamma(A)\, p_A^{(n-1)}(\boldsymbol{x}_A)\, p_{A^c}^{(n-1)}(\boldsymbol{x}_{A^c}) \tag{3.3.13}$$

is the probability a gamete of type $\boldsymbol{x}$ is a member of the gene pool of the population in generation n, where the sum extends over all subsets A of S, A^c is the complement of the set A and $\gamma(A)$ is a linkage probability defined in chapter 2.

Proof: To fix ideas suppose $N = 5$ and consider all genotypes capable of producing the gamete $\boldsymbol{x} = (x_1, x_2, x_3, x_4, x_5)$ in which the alleles x_1 and x_2 were contributed by the maternal parent and the alleles x_3, x_4, and x_5 were contributed by the paternal parent. In the absence of mutation, the probability of a genotype producing this type of gamete is $\gamma(00111)$. If the mating is random, then the probability of a genotype in generation n capable of producing gamete x is

$$p^{(n-1)}(x_1, x_2, y_3, y_4, y_5)\, p^{(n-1)}(y_1, y_2, x_3, x_4, x_5), \tag{3.3.14}$$

where the $y's$ are arbitrary alleles. Therefore,

$$\gamma(00111) p^{(n-1)}(x_1, x_2, y_3, y_4, y_5)\, p^{(n-1)}(y_1, y_2, x_3, x_4, x_5) \tag{3.3.15}$$

is the probability that in generation n the required gamete originates from this particular genotype.

In terms of set notation, the maternal alleles are at the set $A = (1, 2)$ of loci and the paternal alleles are at the complementary set $A^c = (3, 4, 5)$. As $y's$ are arbitrary alleles, to find the desired probability equation (3.3.15) needs to be summed over all $y's$ to obtain

$$\gamma(A) p_A^{(n-1)}(\mathbf{x}_A)\, p_{A^c}^{(n-1)}(\boldsymbol{x}_{A^c}) \,. \tag{3.3.16}$$

For any $N \geq 2$ and subset A of S, a similar argument could be used to derive a probability of the form expressed in (3.3.16). Finally, to obtain the total probability $p^{(n)}(\boldsymbol{x})$ equation (3.3.16) must be summed over all subsets A of S, which completes the proof of the theorem. Observe that the sum in (3.3.13) contains 2^N terms.

It is interesting to note that the result in (3.3.13) could be extended to non-random mating populations by replacing the product on the right in (3.3.15) by a general genotypic distribution in generation $n - 1$. For example, let

$$P^{(n-1)}(x_1, x_2, y_3, y_4, y_5; y_1, y_2, x_3, x_4, x_5)$$

denote the probability a genotype in generation $n - 1$ that may produce a gamete $(x_1, x_2, x_3, x_4, x_5)$ such that alleles x_1 and x_2 are of maternal origin and alleles x_3, x_4, x_5 are of paternal origin. Then, by considering

$$\gamma(00111) P^{(n-1)}(x_1, x_2, y_3, y_4, y_5; y_1, y_2, x_3, x_4, x_5)$$

instead of (3.3.15) the argument could proceed by defining appropriate marginal probabilities similar to those above. No further details will be pursued here but in a subsequent chapter non-random mating systems will be considered.

Equation (3.3.13) is perhaps best understood by considering two special cases. For the two loci case, i.e., $N = 2$, equation (3.3.13) reduces to

$$p^{(n)}(x_1, x_2) = 2\left[\gamma(00)\, p^{(n-1)}(x_1, x_2) + \gamma(10)\, p_1^{(0)}(x_1)\, p_2^{(0)}(x_2)\right]. \tag{3.3.17}$$

In (3.3.17) we have used the result, that under random mating and no selection or mutation, a Hardy-Weinberg equilibrium is attained at each locus in the first generation; namely

$$p_k^{(n)}(x_k) = p_k^{(0)}(x_k) \tag{3.3.18}$$

for all $n \geq 1$ and $k = 1, 2, \ldots, N$. For the case of three loci, equation (3.3.13) takes the form

$$p^{(n)}(\boldsymbol{x}) = 2\begin{bmatrix} \gamma(000)\,p^{(n-1)}(\boldsymbol{x}) \\ +\gamma(100)\,p_1^{(0)}(x_1)\,p_{23}^{(n-1)}(x_2,x_3) \\ +\gamma(010)\,p_2^{(0)}(x_2)\,p_{13}^{(n-1)}(x_1,x_3) \\ +\gamma(110)\,p_{12}^{(n-1)}(x_1,x_2)\,p_3^{(0)}(x_3) \end{bmatrix}, \tag{3.3.19}$$

where $\boldsymbol{x} = (x_1, x_2, x_3)$. It will be noted that (3.3.18) has again been used in the derivation of this equation.

For the case of four loci, equation (3.3.13) contains eight terms of the right-hand side, but, because this equation is difficult to represent on a printed page, no attempt will be made to express this form on the equation here for the case $N = 4$.

In what follows, the limit

$$\lim_{n\uparrow\infty} p^{(n)}(\boldsymbol{x})$$

will be determined for every $N \geq 2$. For the case of two loci, this limit may be easily determined, and to this end let $a = 2\gamma(00)$ and $b = 2\gamma(10)$. Then after one iteration equation (3.3.17) becomes

$$p^{(n)}(x_1,x_2) = a^2 p^{(n-1)}(x_1,x_2) + (1+a)\,bp_1^{(0)}(x_1)\,p_2^{(0)}(x_2).$$

Moreover, by continuing this iterative procedure, it can be shown that

$$p^{(n)}(x_1,x_2) = a^n p^{(0)}(x_1,x_2) + \left(1 + a + \cdots + a^{n-1}\right) bp_1^{(0)}(x_1)\,p_2^{(0)}(x_2). \tag{3.3.20}$$

However,

$$\frac{1-a^n}{1-a} = 1 + a + \cdots + a^{n-1}$$

for $a \neq 1$. Therefore, because $a + b = 1$, equation (3.3.20) reduces to

$$p^{(n)}(x_1,x_2) - p_1^{(0)}(x_1)\,p_2^{(0)}(x_2) = a^n\left(p^{(0)}(x_1,x_2) - p_1^{(0)}(x_1)\,p_2^{(0)}(x_2)\right). \tag{3.3.21}$$

As $0 < a < 1$ by assumption, it follows that

$$\lim_{n\uparrow\infty} p^{(n)}(x_1,x_2) = p_1^{(0)}(x_1)\,p_2^{(0)}(x_2) \tag{3.3.22}$$

for all alleles (x_1, x_2) and the rate of convergence is geometric. In particular, the rate of convergence depends on $a = 2\gamma(00) = 1 - \rho$, which is the probability of no recombination during meiosis, and the deviation

$$\Delta = p^{(0)}(x_1,x_2) - p_1^{(0)}(x_1)\,p_2^{(0)}(x_2). \tag{3.3.23}$$

When the equation holds with respect to two loci, the population is said to be in linkage equilibrium under random mating with no selection or mutation. Observe that if $\Delta = 0$, then initial population is in linkage equilibrium and this equilibrium is maintained in all generation when there is no selection or mutation.

As it turns out, under the conditions of no selection or mutation, the limit in (3.3.22) holds for any number of loci $N \geq 2$. To prove that this indeed the case, write equation (3.3.13) in the form

$$p^{(n+1)}(\boldsymbol{x}) - 2\gamma(00\cdots 0)\, p^{(n)}(\boldsymbol{x}) = \sum_{A} \gamma(A)\, p_A^{(n)}(\boldsymbol{x}_A)\, p_{A^c}^{(n)}(\boldsymbol{x}_{A^c}), \quad (3.3.24)$$

where the sum extends over all subsets A of S except φ, the empty set, and S.

We are thus led to study a different equation of the form

$$c_{n+1} - ac_n = d_n, \quad (3.3.25)$$

where a is a constant such that $0 < a < 1$, $0 < c_n < 1$ and $0 < d_n < 1$ for all $n \geq 1$. In applications of this different equation, it will be assumed that $d_n \to d > 0$ as $n \to \infty$. This convergence also implies that there is a constant c such that $c_n \to c$ as $n \to \infty$. Suppose the sequence c_n does not converge to c. Then the sequence $c_{n+1} - ac_n = d_n$ does not converge, which contradicts the assumption that the sequence d_n does converge to d as $n \to \infty$. Hence, the sequence c_n does converge to c and, moreover,

$$\lim_{n\uparrow\infty} c_n = c = \frac{d}{1-a}. \quad (3.3.26)$$

We are now ready to state and prove another basic result in the theory of random mating populations with no selection or mutation. In the statement that follows, the term arbitrary linkage means that by assumption $0 < 2\gamma(\varphi) < 1$. For all loci $N \geq 2$.

Theorem 3.3.2: If the mating is random and there is no selection or mutation, then for every $N \geq 2$

$$\lim_{n\uparrow\infty} p^{(n)}(\boldsymbol{x}) = \prod_{k=1}^{N} p_k^{(0)}(x_k) \quad (3.3.27)$$

for all gametes $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$ and arbitrary linkage.

Proof: The proof of the theorem is by a principle of mathematical induction that states that if one has a proposition for every integer in the set $(2, 3, 4, \ldots)$, the proposition is true for $N = 2$, and the assumption that it is true of all integers in the set $(2, 3, 4, \ldots, N-1)$ implies it is true for the integer N, then the proposition is true for all the integers in the set $(2, 3, 4, \ldots)$. From the foregoing discussion, it follows that (3.3.27) is true for $N = 2$. By the induction hypothesis suppose (3.3.27) is true of all integers in the set $(2, 3, 4, \ldots, N-1)$. Then, from the right hand side of (3.3.24) it can be seen that

$$\begin{aligned}\lim_{n\uparrow\infty} d_n &= \lim_{n\uparrow\infty} \sum_A \gamma(A)\, p_A^{(n)}(\boldsymbol{x}_A)\, p_{A^c}^{(n)}(\boldsymbol{x}_{A^c}) \\ &= d = (1 - 2\gamma(\varphi)) \prod_{k=1}^{N} p_k^{(0)}(x_k), \qquad (3.3.28)\end{aligned}$$

because the linkage probabilities sum to one. Therefore, from (3.3.28) it follows that

$$\lim_{n\uparrow\infty} p^{(n)}(\boldsymbol{x}) = \prod_{k=1}^{N} p_k^{(0)}(x_k), \qquad (3.3.29)$$

completing the proof of the theorem.

The limiting behavior of the genotypic distribution is an easy consequence of Theorem 3.3.2 and equation (3.3.5) and the details are given in the next theorem.

Theorem 3.3.3: If the mating is random and there is no selection or mutation, then for every $N \geq 2$ and arbitrary linkage

$$\lim_{n\uparrow\infty} P^{(n)}(\boldsymbol{x}, \boldsymbol{y}) = \prod_{k=1}^{N} p_k^{(0)}(x_k)\, p_k^{(0)}(y_k) \qquad (3.3.30)$$

for all choices of gametes $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$ and $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)$, where the symbols x_k and y_k for $k = 1, 2, \ldots, N$ are the elements in the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$.

It is important to notice that, unlike the one locus case, the genotypic distribution is determined by the product of the initial gametic distributions only asymptotically. There is, however, one case in which equation (3.3.21) holds for $n = 1$ and thereafter as long as the mating is random and there is no mutation or selection. The conditions under which this equilibrium in the genotypic distribution is established in the first generation are given in the next theorem.

Theorem 3.3.4: If in the initial generation

$$p^{(0)}(\mathbf{x}) = \prod_{k=1}^{N} p_k^{(0)}(x_k) \tag{3.3.31}$$

for every gamete $\boldsymbol{x}$**,** the mating is random, and there is no mutation or selection, then

$$P^{(n)}(\boldsymbol{x}, \boldsymbol{y}) = \prod_{k=1}^{N} p_k^{(0)}(x_k)\, p_k^{(0)}(y_k) \tag{3.3.32}$$

for every generation $n \geq 1$ and all choices of gametes $\boldsymbol{x}$ and $\boldsymbol{y}$**.**

Proof: Under the hypotheses of the theorem it follows by induction and (3.3.13) that

$$p^{(n)}(\mathbf{x}) = \prod_{k=1}^{N} p_k^{(0)}(x_k) \tag{3.3.33}$$

for all $n \geq 1$ and all choices of gametes $\boldsymbol{x}$**.** The remainder of the theorem follows from equation (3.3.5).

By definition, when (3.3.31) and (3.3.32) hold, the population is said to be in linkage equilibrium with respect to the N loci under consideration, but when these conditions fail to hold, then the population is said to be in linkage disequilibrium. When the number of distinguishable gametes and genotypes is a large number which exceeds the number of individuals in a population, such as that for the HLA-A, HLA-B and HLA-$DRB1$ loci, then is seems plausible that most populations would be in linkage disequilibrium with respect to these loci, simply because the population would not be large enough for all gametes and genotypes to be realized. Therefore, in such populations, the statistical detection of linkage disequilibrium would not necessarily be an indication that natural selection was acting on the alleles at these loci so that an investigator would need to consider other criteria for detecting signatures of natural selection.

When dealing with random mating populations it is frequently assumed that the genotypic distribution is determined by the product of the gametic distributions as is the case in (3.3.32) for every generation $n \geq 1$. We have seen that condition (3.3.32) is, except for special cases, attained only in the limit under random mating and no selection or mutation. It is, therefore, of interest to investigate the rate of convergence to the limit given in Theorem 3.3.2 in the present setting, but such an investigation will not be attempted here.

3.4 Sex Linked Genes in Random Mating Populations

In this section we shall consider sex linked genes in random mating populations with neither selection nor mutation. For the sake of definiteness the female will be regarded as the homogametic sex, i.e., the sex with two homologous X chromosomes, and the male the heterogametic sex, i.e., the sex with a X and a Y chromosome. Such is the case, for example, in man and *Drosophila melanogaster*, the fruit fly, but exactly the opposite is the case in many birds and some insects. That is to say, the female is the heterogametic sex and the male is the homogametic sex. The principles underlying sex linked inheritance are, however, the same in both situations except that the roles of the sexes are interchanged.

By sex linked genes we mean genes located at loci on one arm of the X chromosome. The situation we are going to consider may be represented diagrammatically as

Female	$X\ ----\circ------$	$X\ ----\circ------$
Male	$X\ ----\circ------$	$Y\ ----\circ-$

.

In this table the females is depicted as having two copies of the X chromosome in row 1, but in row 2 the male is depicted as having only one copy of a X chromosome and one copy of a Y chromosome. The symbol $\circ$ stands for the centromere of a chromosome and the dashes $-$ represent segments of a chromosome. In this diagram the so-called sex linked genes are located in the right arm of the X chromosome. It will be noted that the male is haploid with respect to sex linked genes and that he always receives his X chromosome from his mother.

We begin by considering one locus with an arbitrary number of alleles and later on we shall generalize to the case of an arbitrary number of loci with an arbitrary number of alleles at each locus, a pattern which is now becoming familiar. An arbitrary female genotype may be represented in the form (x, y) where the symbol on the left is the allele received from the mother and the symbol on the right is the allele received from the father. Since the male is haploid with respect to sex linked genes, an arbitrary male genotype may be represented by (x). It should be emphasized that the symbols x and y may stand for any of the alleles present at the locus under consideration.

In the large random mating population under consideration let $Q^{(n)}(x, x)$ be the probability of homozygote (x, x) in the female population, and let $P^{(n)}(x, y)$ be the probability of heterozygote (x, y) in the female population in generation $n = 0, 1, 2, \ldots$. When the gametic distributions in the male and female populations differ, then it follows that $P^{(n)}(x, y)$ is not necessarily equal to $P^{(n)}(y, x)$ and it turns out that the situation is easier to handle mathematically if we consider the quantity $2Q^{(n)}(xjy) = P^{(n)}(x, y) + P^{(n)}(y, x)$. The quantity $Q^{(n)}(x, y)$ is the mean of $P^{(n)}(x, y)$ and $P^{(n)}(y, x)$ and $2Q^{(n)}(x, y)$ may be interpreted as the probability that a genotype of the female population carries alleles x and y in generation n. We shall require that the condition

$$\sum_{x} Q^{(n)}(x, x) + 2\sum_{x \neq y} Q^{(n)}(x, y) = 1 \tag{3.4.1}$$

holds for all generations $n \geq 0$. Moreover, if we let $q^{(n)}(x)$ be the probability of gamete x in the gene pool of the female population in generation n, then it follows that with no mutation

$$q^{(n)}(x) = \sum_{y} Q^{(n)}(x, y), \tag{3.4.2}$$

where the sum extends over all alleles present at locus under consideration.

Turning to the male population, we shall let $P^{(x)}(x)$ be the probability of genotype (x) in the male population in generation n. Since the male is haploid with respect to sex linked genes, it is clear that $P^{(n)}(x) = p^{(n)}(x)$, the probability of gamete x in the gene pool of the male population in generation n. Under the assumption of random mating and no selection or mutation, it follows that

$$2Q^{(n+1)}(x, y) = q^{(n)}(x)\, p^{(n)}(y) + q^{(n)}(y)\, p^{(n)}(x) \tag{3.4.3}$$

and since the male always receives his X chromosome from his mother, we have

$$P^{(n+1)}(x) = q^{(n)}(x) \tag{3.4.4}$$

Putting (3.4.3) and (3.4.4) together leads to

$$2Q^{(n+1)}(x, y) = p^{(n+1)}(x)\, p^{(n)}(y) + p^{(n+1)}(y)\, p^{(n)}(x), \tag{3.4.5}$$

which in turn leads to result.

Theorem 3.4.1: If the mating is random and there is no selection or mutation, then the gametic distribution in the male population satisfies the difference equation

$$2p^{(n+2)}(x) = p^{(n+1)}(x) + p^{(n)}(x) \tag{3.4.6}$$

together with the initial conditions $p^{(0)}(x) = p(x)$ and $p^{(1)}\,(x) = q^{(0)}(x) = q\,(x)$. As before x stands for any allele at the locus under consideration.

Proof: We have

$$\begin{aligned} 2p^{(n+2)} &= 2q^{(n+1)}\,(x) \\ &= 2\sum_y Q^{(n+1)}\,(x,y) \\ &= \sum_y \left(p^{(n+1)}\,(x)\,p^{(n)}\,(y) + p^{(n+1)}\,(y)\,p^{(n)}\,(x)\right) \\ &= p^{(n+1)}\,(x) + p^{(n)}\,(x)\,, \end{aligned} \tag{3.4.7}$$

which proves (3.4.6) . The statements, regarding initial conditions, follow from the foregoing discussion.

Difference equation (3.4.6) allows us to easily deduce the limiting form of the gametic distributions in both the male and female populations. The essential features of the situation may be summarized as follows.

Theorem 3.4.2: Set

$$\alpha\,(x) = \frac{p\,(x) + 2q\,(x)}{3} \tag{3.4.8}$$

Then,

$$\lim_{n\uparrow\infty} q^{(n)}(x) = \lim_{n\uparrow\infty} p^{(n)}(x) = \alpha\,(x) \tag{3.4.9}$$

and

$$\lim_{n\uparrow\infty} Q^{(n)}\,(x,y) = \alpha\,(x)\,\alpha\,(y) \tag{3.4.10}$$

for all choices of alleles x and y.

Proof: The proof of the theorem is based on the solution of difference equation (3.4.6). To find the solution of this equation suppose

$$p^{(n)}\,(x) = \beta^n\,(x)\,.$$

Substituting this equation into (3.4.6) leads to the quadratic equation

$$2\beta^2\,(x) - \beta\,(x) - 1 = 0$$

with roots $\beta_1(x) = 1$ and $\beta_2(x) = -1/2$. From this result it is clear that $\beta_1(x)$ and $\beta_2(x)$ do not depend on the allele x. Moreover, since $p^{(n)}(x) = 1$

and $p^{(n)}(x) = (-1/2)^n$ are solutions of (3.4.6), it follows that the general solution has the form

$$p^{(n)}(x) = c_1(x) + c_2(x)\left(-\frac{1}{2}\right)^n,$$

where $c_1(x)$ and $c_2(x)$ are to be determined.

From the initial conditions, it follows that

$$\begin{aligned} c_1(x) + c_2(x) &= p(x) \\ c_1(x) - \frac{c_2(x)}{2} &= q(x). \end{aligned}$$

Therefore,

$$c_1(x) = \frac{p(x) + 2q(x)}{3} = \alpha(x)$$

and

$$c_2(x) = \frac{2(p(x) - q(x))}{3}.$$

Therefore,

$$p^{(n)}(x) = \alpha(x) + \frac{2(p(x) - q(x))}{3}\left(-\frac{1}{2}\right)^n \tag{3.4.11}$$

is the unique solution of difference equation (3.4.6) satisfying the initial conditions $p^{(0)}(x) = p(x)$ and $p^{(1)}(x) = q(x)$. Equations (3.4.9) and (3.4.10) now easily follow by letting $n \uparrow \infty$ in (3.4.11) as well as in (3.4.4) and (3.4.5).

It is of interest to note that the sequence $(p^{(n)}(x))$ approaches its limit $\alpha(x)$ in an oscillatory manner in the sense that the difference $p^{(n)}(x) - \alpha(x)$ changes sign in every generation. Evidently, the rate of convergence to the limit $\alpha(x)$ will in all cases be quite rapid, but will depend to a large extent on the value of $\mid p(x) - q(x) \mid$. For the sake of simplicity assume $\mid p(x) - q(x) \mid \leq 1/2$. Then $\mid p^{(n)}(x) - \alpha(x) \mid \leq \frac{1}{3}\left(\frac{1}{2}\right)^n = 8.138\,020\,833\,333\,33 \times 10^{-5}$

$$\mid p^{(n)}(x) - \alpha(x) \mid \leq \frac{1}{3}\left(\frac{1}{2}\right)^n \leq 8.138 \times 10^{-5}$$

if $n \geq 12$. In other words, $p^{(n)}(x)$ and $\alpha(x)$ will, in this case, differ approximately by at most 8.138×10^{-5} after about twelve generations.

The results obtained thus far provide an easy stepping-stone to the elucidation of the situation for the case of $N \geq 2$ loci with arbitrary linkage and with an arbitrary number of alleles at each locus. Just as in the one locus case an arbitrary female genotype will be represented by the ordered pair of vectors $(\boldsymbol{x}, \boldsymbol{y})$, where $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$ and $y = (y_1, y_2, \ldots, y_N)$ are the alleles received from the maternal and paternal parents, respectively. By convention the symbol on the left in the ordered pair $(\boldsymbol{x}, \boldsymbol{y})$ will always represent the genes contributed by the maternal parent. In analogy with the one locus case an arbitrary male genotype will be represented by $(\boldsymbol{x})$, where $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$.

For the case of $N \geq 2$ loci we shall let $Q^{(n)}(\boldsymbol{x}, \boldsymbol{y})$ be the probability of genotype $(\boldsymbol{x}, \boldsymbol{y})$ in the female population and $q^{(n)}(\boldsymbol{x})$ the probability of gametic $\boldsymbol{x}$ in the gene pool of the female population in generation n. Similarly, let $P^{(n)}(\boldsymbol{x})$ be the probability of genotype $(\boldsymbol{x})$ in the male population in generation n. Since the male is haploid with respect to sex linked genes $P^{(n)}(\boldsymbol{x}) = p^{(n)}(\boldsymbol{x})$, the probability of gamete $\boldsymbol{x}$ in the gene pool of the male population in generation n, just as it was in the one locus case.

Under random mating and no selection,

$$Q^{(n+1)}(\boldsymbol{x}, \boldsymbol{y}) = q^{(n)}(\boldsymbol{x}) p^{(n)}(\boldsymbol{y}) \tag{3.4.12}$$

and because a male always receives his sex linked genes from his mother

$$P^{(n+1)}(\boldsymbol{x}) = p^{(n+1)}(\boldsymbol{x}) = q^{(n)}(\boldsymbol{x}) \tag{3.4.13}$$

for all $n \geq 0$.

Our principal objective in the remainder of this section is to determine the limiting form of the gametic distribution in both the male and female populations. From what precedes we should expect two things. Firstly, the gametic distributions in the male and female populations should be equalized in the limit, and secondly, the limiting gametic distribution should be completely determined by the product of the marginal gametic distributions associated with each locus, i.e., independence is attained in the limit. As we shall see, the situation just described is indeed the case.

Due to the fact that crossing over with respect to sex linked genes occurs only in the female, our first task will be that of finding a system of difference equations satisfied by the gametic distributions of the female population in successive generations. With each subset A of the set $S = (1, 2, \ldots, N)$ we may associate the marginal probabilities $p_A^{(n)}(\boldsymbol{x}_A)$ and $q_{A^c}^{(n)}(\mathbf{x}_{A^c})$ for $n = 0, 1, 2, \ldots$. These marginal probabilities are defined in the same way as they were in section 3.3. The desired system of difference equations is the subject matter of the next theorem.

Theorem 3.4.3: If the mating is random and there is no mutation or selection, then

$$q^{(n+1)}(\boldsymbol{x}) = \sum_{A} \gamma\left(A\right) q_A^{(n)}(\boldsymbol{x}_A) q_{A^c}^{(n-1)}\left(\mathrm{x}_{A^c}\right) \tag{3.4.14}$$

for every generation $n \geq 1$ and all choices of gametes $\boldsymbol{x}$. The sum in (3.4.14) extends over all subsets of S and $\gamma(A)$ is the linkage probability defined before.

Proof: The proof of Theorem 3.4.3 follows along the same lines as the proof of Theorem 3.3.1. Briefly, if the mating is random and there is no mutation or selection, then

$$\gamma\left(A\right) q_A^{(n)}(\boldsymbol{x}_A) p_{A^c}^{(n)}\left(\mathrm{x}_{A^c}\right) = \gamma\left(A\right) q_A^{(n)}(\boldsymbol{x}_A) q_{A^c}^{(n-1)}\left(\mathrm{x}_{A^c}\right) \tag{3.4.15}$$

is the probability that in generation $n+1$ gamete $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$ is produced in the population with maternal genes at the loci of the set A and paternal genes at the loci of the set A^c. The theorem follows by summing over all subsets of S.

At each locus under consideration the results of Theorem 3.4.2 hold. Set

$$\lim_{n\uparrow\infty} q_k^{(n)}(x_k) = \lim_{n\uparrow\infty} p_k^{(n)}(x_k) = \alpha_k\left(x_x\right)$$

for $k = 1, 2, \ldots, N$. The principal limit theorem for the case of $N \geq 2$ sex linked loci with arbitrary linkage and an arbitrary number of alleles at each locus follows easily from Theorem 3.4.3.

Theorem 3.4.4: If the mating is random and there is no selection or mutation, then for every $N \geq 2$ and arbitrary linkage

$$\lim_{n\uparrow\infty} q^{(n)}(\boldsymbol{x}) = \lim_{n\uparrow\infty} p^{(n)}(\boldsymbol{x}) = \prod_{k=1}^{N} \alpha_k\left(x_x\right) \tag{3.4.16}$$

and

$$\lim_{n\uparrow\infty} Q^{(n)}(\boldsymbol{x}, \boldsymbol{y}) = \prod_{k=1}^{N} \alpha_k\left(x_x\right) \alpha_k\left(y_k\right) \tag{3.4.17}$$

for all choices of gametes $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$ and $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)$.

Proof: The proof is by the principle of mathematical induction used in the proof of Theorem 3.3.2. For the two locus case, let equation (3.4.14) becomes

$$q^{(n+1)}(x_1,x_2) = \gamma(\varphi)\begin{pmatrix} q^{(n)}(x_1,x_2) \\ +q^{(n-1)}(x_1,x_2) \end{pmatrix} \tag{3.4.18}$$
$$+\gamma(10)\begin{pmatrix} q_1^{(n)}(x_1)\,q_2^{(n-1)}(x_2) \\ +q_1^{(n-1)}(x_1)\,q_2^{(n)}(x_2) \end{pmatrix}.$$

To simplify the writing of this equation, let $a_n = q^{(n)}(x_1,x_2)$ and let

$$d_n = q_1^{(n)}(x_1)\,q_2^{(n-1)}(x_2) + q_1^{(n-1)}(x_1)\,q_2^{(n)}(x_2).$$

Then, equation (3.3.18) takes the simple form

$$a_{n+1} - \gamma(\varphi)(a_n + a_{n-1}) = \gamma(10)\,d_n.$$

Next observe that

$$\lim_{n\uparrow\infty}\gamma(10)\,d_n = 2\gamma(10)\,\alpha_1(x_1)\,\alpha_2(x_2).$$

Therefore, if we proceed as in the proof of Theorem 3.3.2, it follows that the sequence a_n converges to a limit so that

$$\lim_{n\uparrow\infty} q^{(n)}(x_1,x_2) = \frac{2\gamma(10)}{1-2\gamma(\varphi)}\alpha_1(x_1)\,\alpha_2(x_2) = \alpha_1(x_1)\,\alpha_2(x_2).$$

This proves the theorem for the case $N = 2$.

That statement (3.4.16) holds for arbitrary $N \geq 2$ follows by writing (3.4.14) in the form

$$q^{(n+1)}(\boldsymbol{x}) - \gamma(\varphi)\left(q^{(n)}(\boldsymbol{x}) + q^{(n-1)}(\boldsymbol{x})\right) = \sum_A \gamma(A)\,q_A^{(n)}(\boldsymbol{x}_A)q_{A^c}^{(n-1)}(\mathbf{x}_{A^c})$$

where the sum extends over all subsets A of S except φ and S, and proceeding as in the induction proof of Theorem 3.3.2. Equation (3.4.17) follows by letting $n \to \infty$ in (3.4.12). This completes the proof of the theorem.

We shall close this section by pointing out some examples of sex linked inheritance. Perhaps two of the most famous examples of sex linked inheritance in man are color blindness and hemophilia or bleeders disease. Hemophilia is particularly famous due to its prevalence among the males of the royal houses of Europe during the nineteenth century. Both these abnormal traits are inherited as recessives in the female, but, since males carry only one copy of sex linked gene, a recessive gene always manifests itself in the male. Recombination between color blindness and hemophilia has been observed so that we may infer that these genes are situated at different loci on the X chromosome. For a presentation of the evidence regarding these statements and a thorough discussion of sex linked inheritance in man the reader should consult Stern (1960). Examples of sex linked inheritance in other organisms may be found in almost any book on classical genetics such as Fristrom and Spieth (1980). For instance, many examples of sex linked genes are known in the fruit fly, *Drosophila melanonaster*, the organism from which a large part of our knowledge of classical genetics stems.

It is perhaps somewhat ironical, but by no means discouraging, that the multilocus model of sex linked genes discussed above does not seem to be applicable to color blindness and hemophilia in man for at least two reasons. Firstly, males possessing the recessive gene for hemophilia seem to be selected against in the sense that they frequently die before they contribute offspring to the population. And secondly, mating with respect to the hemophilic locus may not be random from the point of view of the population as a whole, because bleeders seem to occur within pockets of the population which are composed of people more highly related than those of the general population. These pockets may be due to geographical isolation such as mountain ranges or to social isolation such as that which occurs in royal families. For further details see Stern (1960).

3.5 Comments and Historical Notes

Although Theorem 3.2.1 is not always stated in the form given here, its subject matter has been known since about 1908. The Hardy-Weinberg law was discovered independently by Hardy (1908), a well-known British mathematician, and Weinberg (1908), a German physician. The terminology, The Hardy-Weinberg Law, is due to Stern who translated an important section of Weinberg's paper in Stern (1943). The extension of the Hardy-Weinberg law to the case of multiple alleles at one locus is an adaptation of a result which may be found in Kempthorne (1957).

Section 3.3 is an adaptation of the work of Geirenger (1944). It should also be mentioned that the notation of the entire chapter is patterned after that of Geirenger, who deserves a large measure of the credit for the introduction of set theoretic methods which permit us to handle the case of an arbitrary number of loci with ease. Theorem 3.3.2 for the special case of two loci seems to have been known as early as (1917) by Jennings (1917) and (1923) and also by Robbins (1918a) and (1918b). Theorems 3.4.1 and 3.4.2 are patterned after some results of Kempthorne (1957) chapter two, but Theorems 3.4.3 and 3.4.4, which are easy extensions of the results of section 3.3, seem to be new.

The present chapter was written in a language describing two sex populations, but the results, with the exception of those in section 3, are thought to be applicable to populations of annual plants which under uncontrolled conditions cross fertilize. It generally is agreed that many plant populations approximate random mating in the sense that the truth of condition 3.3.3 is plausible. An example of such a plant is *Zea mays* or common field corn.

An extension of the results of this chapter to the case of overlapping generations would be most interesting and timely. For a simple deterministic extension of the Hardy-Weinberg Law to the case of over lapping generations with two alleles at a locus see Moran (1962), pages 23 and 24. The Hardy-Weinberg Law may also be deduced within the framework of a multitype Galton-Watson process with discrete time generations. A discussion of this extension may be found in Mode (1971) on pages 7 and 8. Furthermore, a discussion of a derivation of the Hardy-Weinberg Law within the framework of generalized multitype branching processes, which accommodate over lapping generations, may be found on page 130 of Mode (1971). It should be mentioned that within the framework of branching processes attention is focused on the evolution of a sub-population within a large population arising from a single progenitor but not a population as a whole. Therefore, derivations of the Hardy-Weinberg Law within the framework of branching processes should be viewed as a preliminary results, with the hope that in time it may be possible to deduce this law within a more general framework of stochastic genetic evolutionary processes that accommodates mating among genotypes.

Bibliography

[1] Buerger, R. (2000) **The Mathematical Theory of Selection, Recombination and Mutation**. John Wiley & Sons, LTD. Chichester, New York, Weinheim, Brisbane, Singapore, Toronto.

[2] Christiansen, F. B. (2000) **Population Genetics of Multiple Loci**. John Wiley & Sons, LTD. Chichester, New York, Weinheim, Brisbane, Singapore, Toronto.

[3] Crow, J. F. and Kimura, M. (1970) **An Introduction to Population Genetics Theory.** Harper & Row, New York, Evenston, and London.

[4] Cavalli-Sforza, L. L. and Bodmer, W. F. (1971) **The Genetics of Human Populations**. W. H. Freeman and Company, San Francisco.

[5] Ewens, W. J. (2004) **Mathematical Population Genetics I. Theoretical Introduction**, second edition. Springer, New York, Heidelberg, London, Paris, Tokyo.

[6] Fristrom, J. W. and Spieth, P. T. (1980) **Principles of Genetics**. Chiron Press, Blackwell Scientific Publication, Oxford, London, Edinburgh, Melbourne.

[7] Geirenger, H. (1944) On the Probability Theory of Linkage in Mendelian Heredity. Ann. Math. Stat. **15**:25–57.

[8] Hardy, G. H. (1908) Mendelian Proportions in a Mixed Population. Science **20**:49–50.

[9] Hinds D., Stuve L., Nilsen G., Halperin E., Eskin E., Ballinger D., Kelly K., Frazer A., and Cox D. (2005) Whole-genome pattern of common DNA variation in three human populations. Science **307**:1072–1079.

[10] Jennings, H. S. (1917) The Numerical Results of Diverse Systems of Breeding with Respect to Two Pairs of Characters. Genetics **12**:97–154.

[11] Jennings, H. S. (1923) The numerical Relations in the Crossing Over of the Genes with a Critical Examination of the Theory That the Genes are Arranged in a Linear Series. Genetics **8**:393.

[12] Kempthorne, O. (1957) **An Introduction to Genetic Statistics**. John Wiley and Sons, Inc., New York.

[13] Mode, C. J. (1971) **Multitype Branching Processes-Theory and Applications**. Elsevier, New York, London, Amsterdam.

[14] Moran, P. A. P. (1962). **The Statistical Processes of Evolutionary Theory**. Clarendon Press, Oxford.

[15] Robbins, R. B. (1918a) Applications of Mathematics to Breeding Problems II. Genetics **3**:73–92.

[16] Robbins, R. B. (1918b) Applications of Mathematics to Breeding Problems III. Genetics **3**:375–389.

[17] Stern, C. (1943) The Hardy-Weinberg Law. Science **97**:137–138.

[18] Stern, C. (1960) **Principles of Human Genetics, Second edition**. W. H. Freeman and Company, San Francisco, California.

[19] Weinberg, W. (1908) Uber den Naehweis der Verenbung bein Menschen. Jahreschefte Verein f. veterl. Naturk. in Wurttenberg. **64**:368–382.

Chapter 4

Two Allele Wright-Fisher Process with Mutation and Selection

4.1 Introduction

During the last seven to ten decades, Wright-Fisher models have been given a considerable amount of attention in theoretical population genetics. Briefly, Wright-Fisher models are gamete sampling models in the sense that rather than focusing on the genotypes and phenotypes as the basic elements of the theory, attention has been focused on finite samples of gametes or haplotypes as they evolve from generation to generation. Thus, if a population of diploids consists of N individuals in some generation and population size is restricted so that the next generation also consists of N individuals, then in any generation there are $2N$ gametes or haplotypes to consider. From the probabilistic point of view, the evolution of such a population over many generations, would consists of sampling $2N$ gametes from N individuals with replacement from a given generation to produce the gametes of the next generation. The assumption of random sampling from generation to generation leads to consideration of arrays of binomial distributions as will be illustrated in the sections that follow in this chapter.

One of the implicit assumptions underlying this sampling process from generation to generation is that the sample in any generation depends only on the population of gametes in the preceding generation, which implies that the type of stochastic process under consideration may be formulated as a Markov chain. In such a stochastic process, the evolution of the process among a set of states depends only on the state occupied in the immediate past. In the next section of this chapter, a brief overview of the theory of Markov chains will be given and in subsequent sections the properties of Wright-Fisher processes, formulated as Markov chains, will be explored in detail for the case of an autosomal locus with two alleles. It should also be mentioned that there is a rather large literature on Wright-Fisher processes and much of this literature is devoted to diffusion approximation to the process, see for example, the book by Ewens (2004). In this and later chapters in this book, however, attention will be devoted to computational methods which seem to be appropriate for the present age of powerful desk top computers on which one can do matrix computations, using such software packages as MATLAB, APL 2000 and S-PLUS. But from time to time, relevant results from the theory of diffusion processes will be mentioned.

4.2 Overview of Markov Chains with Stationary Transition Probabilities

Markov chains with stationary transition probabilities are a class of stochastic processes that evolve in discrete time, and have been applied in various forms in mathematical population genetics. To formulate a model as a Markov chain in discrete time, one needs to specify a set $\mathfrak{S}$ of states such that the process evolves among the elements of $\mathfrak{S}$. Elements of $\mathfrak{S}$ will be denoted by the symbols $i, j, k, \ldots$ with or without subscripts. Let X_n for $n = 0, 1, 2, \ldots$ denote a sequence of random variables, taking values in the set $\mathfrak{S}$ that denote the state of the process at time n. The event that the process is in state $j \in \mathfrak{S}$ at time n will be denoted by $[X_n = j]$ In genetics, these discrete times are often thought of as generations of some species under consideration. The event that the process was in state i_ν at times $\nu = 0, 1, 2, \ldots, n$ will be denoted by the symbol

$$[X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n] \tag{4.2.1}$$

for every $n \geq 1$.

To formally describe a model as a stochastic process, it is necessary to define all the finite dimensional distributions of the process. In principle, these distributions may be defined in terms of conditional probabilities. To illustrate this idea, consider the event in (4.2.1) for the case $n = 2$. Then, given that $X_0 = i_0$,

$$\begin{aligned} &P\left[X_1 = i_1, X_2 = i_2 \mid X_0 = i_0\right] \\ &= P\left[X_1 = i_1 \mid X_0 = i_0\right] P\left[X_2 = i_2 \mid X_0 = i_0, X_1 = i_1\right]. \end{aligned} \tag{4.2.2}$$

An assumption that characterizes Markov chains in discrete time is that a probability of a transition from one state to another, depends only on the most recent state visited by the process. Under this assumption, the probability on the right in (4.2.2) takes the form

$$P\left[X_2 = i_2 \mid X_0 = i_0, X_1 = i_1\right] = P\left[X_2 = i_2 \mid X_1 = i_1\right]. \tag{4.2.3}$$

In general, to formulate a model as a Markov chain, it will be assumed that for any $n \geq 1$

$$P\left[X_n = i_n \mid X_{n-1} = i_{n-1}, \ldots, X_0 = i_0\right] = P\left[X_n = i_n \mid X_{n-1} = i_{n-1}\right] \tag{4.2.4}$$

for all states $i_0, i_1, \ldots, i_{n-1}$. This assumption is also known as the Markov property.

A Markov chain is said to have stationary transition probabilities if for every pair i, j of states in $\mathfrak{S}$ and $n \geq 1$

$$P\left[X_n = j \mid X_{n-1} = i\right] = p_{ij}, \tag{4.2.5}$$

where p_{ij} is a constant. Observe that, according to the assumption of stationary transition probabilities, the probability on the left does not depend on $n \geq 1$. The matrix of conditional probabilities

$$\boldsymbol{P} = \left(p_{ij} \mid i \in \mathfrak{S}, j \in \mathfrak{S}\right) \tag{4.2.6}$$

is known as the transition matrix of the Markov chain. This matrix has the properties that $p_{ij} \geq 0$ for all pairs i, j and for every $i \in \mathfrak{S}$,

$$\sum_{j \in \mathfrak{S}} p_{ij} = 1. \tag{4.2.7}$$

Given the transition matrix of the process, all finite dimensional distributions of the process are defined by equations of the form

$$P\left[X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n\right] = p_{i_0} \prod_{\nu=1}^{n} p_{i_{\nu-1} i_\nu} \tag{4.2.8}$$

for all states $i_0, i_1, \ldots, i_n$ in $\mathfrak{S}$ and $n \geq 1$, where $\left(p_{i_0} \mid i_0 \in \mathfrak{S}\right)$ is an assigned initial distribution.

Many probabilities of interest for a Markov chain with stationary transition probabilities may be computed in terms of powers of the transition matrix $\boldsymbol{P}$. For example, consider the conditional probability

$$P\left[X_2 = j \mid X_0 = i\right]. \tag{4.2.9}$$

If a process is in state j at time 2, given that the initial state was i, then the first transition out of i was to some state k with probability p_{ik}. Given that the process was in state k at time 1, the conditional probability of the transition to state j is p_{kj} and, given this path of states and the Markov property, $p_{ik}p_{kj}$ is the probability of the process is in state j at step or time 2. By summing over all states $k \in \mathfrak{S}$, it follows that

$$P\left[X_2 = j \mid X_0 = i\right] = \sum_{k \in \mathfrak{S}} p_{ik}p_{kj} = p_{ij}^{(2)}, \tag{4.2.10}$$

where, by definition, $p_{ij}^{(2)}$ is element i, j in the matrix $\boldsymbol{P}^2$. In general, for every $n \geq 1$, let $p_{ik}^{(n)}$ be the element i, k in the matrix $\boldsymbol{P}^n$, the *n-th* power of the matrix $\boldsymbol{P}$. Then, by using an argument similar to that in (4.2.10), it follows that

$$P\left[X_n = j \mid X_0 = i\right] = \sum_{k \in \mathfrak{S}} p_{ik}^{(n-1)}p_{kj} = p_{ij}^{(n)}, \tag{4.2.11}$$

where $p_{ij}^{(n)}$ is element i, j in the matrix $\boldsymbol{P}^n$, the *n-th* power of the matrix $\boldsymbol{P}$. This equation will be valid for all $n \geq 1$ if $\boldsymbol{P}^0$ is defined by $\boldsymbol{P}^0 = \boldsymbol{I}$, an identity matrix, and $\boldsymbol{P}^1 = \boldsymbol{P}$.

It is, of course, possible to define a probability space $(\Omega, \mathcal{A}, P)$ underlying the Markov chain just described, which depends only on the state space $\mathfrak{S}$ and the transition matrix (p_{ij}). But, for the time being, the details involving the construction of this space will not be discussed and attention will be focused on special cases of the structure defined in this section that have been used extensively in population genetics and will be used in the subsequent sections of this chapter.

4.3 Overview of Wright-Fisher Perspective

From a cursory inspection of the literature, it is not entirely clear whether Wright or Fisher invented the type of model to be discussed in this section. Karlin and Taylor (1975), see page 56, discuss this model as an example of a Markov chain with a finite state space and mention the famous book by Fisher (1958) in a foot note. Rather than pursuing the origins of the term here, which is best left to some future historian of science, we will proceed directly to set forth the mathematical structure of the model and discuss its genetic ramifications. The treatment that follows will provide a general overview that seems appropriate during an age of increasing technological advances in powers and usefulness of computers and the sequencing of human genome as well as those of other species.

Consider a population of some finite size $N \geq 1$ of diploid individuals with respect to two autosomal alleles A_1 and A_2 at one locus. Because each individual in the diploid population has two genes at this locus, in any generation of a population of N individuals there are $2N$ genes. Let the random variable X_n denote the number A_1 genes in the population in generation $n = 0, 1, 2, \ldots$. Then, because, by assumption, population size is constant from generation to generation, the number of A_2 genes in the population in generation n is $2N - X_n$. The range of the random variable X_n is the set

$$\mathfrak{S} = \{i \mid i = 0, 1, 2, \ldots, 2N\} \tag{4.3.1}$$

for all generations n. Note, there are $2N + 1$ elements in the set $\mathfrak{S}$.

The set $\mathfrak{S}$ will also be the state space of a discrete time parameter Markov chain which is defined as follows. For any generation $n \geq 1$, let

$$P\left[X_n = j \mid X_{n-1} = i\right] = p_{ij} \tag{4.3.2}$$

be the conditional probability that if the population is in state $i \in \mathfrak{S}$ in generation $n-1$, there is a transition to state $j \in \mathfrak{S}$ in generation n. It will be assumed that this probability does not depend on n. In Wright-Fisher models, it is assumed that this transition probability has the form

$$p_{ij} = \binom{2N}{j}\left(\frac{i}{2N}\right)^j\left(1 - \frac{i}{2N}\right)^{2N-j} \tag{4.3.3}$$

for all i, $j = 0, 1, 2, \ldots, 2N$. It is easy to recognize this formula as that for the probability density function ($p.d.f.$) of the binomial distribution. From the probabilistic point of view, this formula may be interpreted as sampling a population of gametes at random with replacement until a sample of size $2N$ is obtained.

The argument that is usually given to justify this formulation from the genetic point of view is that if the mating is random among the diploid individuals in each generation and there is no mutation or selection, then from the mathematical point of view, it suffices to consider the frequencies of the alleles A_1 and A_2 in the population in any generation. Thus, if $X_{n-1} = i$, then the frequency of A_1 in the population in generation $n - 1$ is

$$p_i = \frac{i}{2N} \tag{4.3.4}$$

and that for A_2 is

$$q_i = 1 - \frac{i}{2N}. \tag{4.3.5}$$

This type of model, which is sometimes referred as one member of a class of chain binomials models, has pertinent implications to the effect that somehow the randomness and matings and the assortment of genes in the population act is such a way that individual genes are conditionally independent in the sampling process from generation to generation in the following sense. Let

$$(\boldsymbol{\xi}_\nu \mid \nu = 1, 2, \ldots, 2N) \tag{4.3.6}$$

be a set of conditionally independent Bernoulli indicators such that if $X_{n-1} = i$, then

$$P\left[\boldsymbol{\xi}_\nu = 1\right] = p_i \tag{4.3.7}$$

and

$$P\left[\boldsymbol{\xi}_\nu = 0\right] = 1 - p_i = q_i. \tag{4.3.8}$$

Then, given that $X_{n-1} = i$, the random variable X_n has the representation

$$X_n = \sum_{\nu=1}^{2N} \boldsymbol{\xi}_\nu \tag{4.3.9}$$

for $n \geq 1$. This observation is included here to emphasize that in many population models formulated within a stochastic paradigm, the idea of conditional independence among individuals is frequently used.

From this representation of X_n, it is easy to compute the conditional expectation on X_n, given that $X_n = i$. It is

$$E\left[X_n \mid X_{n-1} = i\right] = 2Np_i = i = X_{n-1}, \tag{4.3.10}$$

and the conditional variance is

$$var\left[X_n \mid X_{n-1} = i\right] = 2Np_i(1 - p_i) \ = i\left(1 - \frac{i}{2N}\right). \tag{4.3.11}$$

Equation (4.3.10) implies a very interesting result that follows by taking unconditional expectations. For example, the unconditional expectation of X_n is

$$E\left[X_n\right] = E\left[E\left[X_n \mid X_{n-1}\right]\right] = E\left[X_{n-1}\right] \tag{4.3.12}$$

for all $n \geq 1$. In particular, suppose $X_0 = i \in \mathfrak{S}$. Then,

$$E\left[X_1\right] = i, \tag{4.3.13}$$

and by induction it can be shown that

$$E\left[X_n\right] = i \tag{4.3.14}$$

for all $n \geq 1$.

Without going into the technical details, equation (4.3.10), along with the Markov property, also implies that the sequence $(X_n \mid n \geq 0)$ is a martingale, see Karlin and Taylor (1975) for details. As (4.3.14) is finite for all n, by a well known theorem from martingale theory, it follows that there is a random variable X such that

$$X_n \to X \tag{4.3.15}$$

as $n \uparrow \infty$ with probability one. However, this theoretical result is not particularly interesting unless it is possible to say something about the distribution of X. Later on this distribution will be described in detail. Because all expectations in the Wright-Fisher model are finite sums, it follows that

$$\lim_{n\uparrow\infty} E\left[X_n \mid X_0 = i\right] = E\left[\lim_{n\uparrow\infty} X_n \mid X_0 = i\right] = E\left[X \mid X_0 = i\right] = i \tag{4.3.16}$$

for all $i \in \mathfrak{S}_2$. As will be demonstrated subsequently, this result will be very useful in deducing for the probability that an allele becomes fixed eventually in a population.

4.4 Absorbing Markov Chains with a Finite State Space

From the mathematical perspective the Wright-Fisher model under consideration belongs to the class of so-called absorbing Markov chains with a finite state space. A state is said to be absorbing if the process remains there for all time, giving that this state was entered. Let $i \in \mathfrak{S}$ denote an absorbing state. Then, by definition $p_{ii} = 1$ and $p_{ij} = 0$ if $i \neq j$. For the Wright-Fisher model under consideration, the states $i = 0$ and $i = 2N$ are absorbing. For if $i = 0$, then all genes in the population of N individuals carry are copies of gene A_2. They are all, therefore, homozygous with genotype A_2A_2. Similarly, if $i = 2N$, then all individuals in the population are homozygous with genotype A_1A_1. It is assumed, of course, that the two alleles under consideration do not mutate.

There is a canonical way of viewing absorbing Markov chains with a finite state space, due to Kemeny and Snell (1976). Unlike many mathematicians of his day, John Kemeny was interested in computer science and was also instrumental in designing the BASIC programming language, a language that is used widely today. This interest seems to be reflected in the developments that follow.

Consider a Markov transition matrix

$$\mathbf{P} = (p_{ij}) \tag{4.4.1}$$

and suppose there are $r_1 \geq 1$ absorbing states and $r_2 \geq 1$ transient states for a total of $r = r_1 + r_2$ states. By a transient state we mean roughly a state such that a stochastic process will leave it eventually with probability one. A useful way of representing the transition matrix is the partitioned form

$$\mathbf{P} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{R} & \mathbf{Q} \end{bmatrix}, \tag{4.4.2}$$

where $\mathbf{I}$ is a $r_1 \times r_1$ identity matrix, $\mathbf{0}$ is a $r_1 \times r_2$ matrix of zeros, $\mathbf{R}$ is a $r_2 \times r_1$ matrix of transition probabilities governing transitions for the set of transient states to the set of absorbing states and $\mathbf{Q}$ is a $r_2 \times r_2$ governing transitions among transient states.

For the Wright-Fisher model under consideration, as mentioned above, the set of absorbing states is

$$\mathfrak{S}_1 = (0, 2N) \tag{4.4.3}$$

so that $r_1 = 2$ and the set of transient states is

$$\mathfrak{S}_2 = \{i \mid i = 1, 2, \ldots, 2N - 1\}. \tag{4.4.4}$$

Observe that the number of states in the set $\mathfrak{S}_2$ is $r_2 = 2N - 1$. Furthermore, for this model the matrix $\mathbf{R}$ is $(2N-1) \times 2$ and for every $i \in \mathfrak{S}_2$, the $i - th$ row of this matrix has the form

$$\left(\binom{2N}{0} p_i^0 q_i^{2N-0}, \binom{2N}{2N} p_i^{2N} q_i^{2N-2N} \right) = \left(q_i^{2N}, p_i^{2N} \right). \tag{4.4.5}$$

When N is large, both these probabilities will be small. Finally, the $(2N-1)\times$ $(2N-1)$ matrix $\mathbf{Q}$ has the form

$$\mathbf{Q} = (p_{ij}) = \left(\binom{2N}{j} p_i^j q_i^{2N-j} \mid i \in \mathfrak{S}_2, j \in \mathfrak{S}_2 \right). \tag{4.4.6}$$

If one had to do all the symbolic calculations with just a pencil and paper, the task of constructing these matrices would be formidable. However, it would not be difficult to write a computer program to compute the entries in these matrices as a function of $N \geq 1$.

One of the results of interest for the model under consideration is that of finding a numerical or symbolic version of the $r_2 \times r_1$ matrix conditional probabilities

$$\mathbf{F}^{(n)} = \left(f_{ij}^{(n)} \right) = \left(P\left[X_n = j \mid X_0 = i\right] \mid i \in \mathfrak{S}_2, j \in \mathfrak{S}_1 \right). \tag{4.4.7}$$

Observe that each element of this matrix is the conditional probability that the process reaches absorbing state j by generation n, given that the initial transient state was i. In terms of a Wright-Fisher model, observe that if $j = 0$, then the gene A_2 becomes fixed in the population and, if $j = 2N$, then the gene A_1 is fixed in the population. Consequently, the absorption probabilities in the matrix $\mathbf{F}^{(n)}$ are sometimes referred as the probabilities of fixation by generation n.

To derive a symbolic formula for this matrix, the transition matrix $\mathbf{P}$ needs to be raised to the *n-th* power. By using the symbolic computation engine connected with this word processes, it can be shown that

$$\mathbf{P}^6 = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \left(\mathbf{I}+\mathbf{Q}+\mathbf{Q}^2+\cdots+\mathbf{Q}^5\right)\mathbf{R} & \mathbf{Q}^6 \end{bmatrix}. \tag{4.4.8}$$

And by induction it can be shown that

$$\mathbf{P}^n = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \left(\mathbf{I}+\mathbf{Q}+\mathbf{Q}^2+\cdots+\mathbf{Q}^{n-1}\right)\mathbf{R} & \mathbf{Q}^n \end{bmatrix} \tag{4.4.9}$$

for every $n \geq 2$. Hence, a general formula for the matrix of absorption probabilities is

$$\mathbf{F}^{(n)} = \left(\mathbf{I}+\mathbf{Q}+\mathbf{Q}^2+\cdots+\mathbf{Q}^{n-1}\right)\mathbf{R}. \tag{4.4.10}$$

But,

$$\left(\mathbf{I}+\mathbf{Q}+\mathbf{Q}^2+\cdots+\mathbf{Q}^{n-1}\right)(\mathbf{I}-\mathbf{Q}) = \mathbf{I}-\mathbf{Q}^n. \tag{4.4.11}$$

So if $\mathbf{Q}^n \longrightarrow \mathbf{0}$, a zero matrix as $n \uparrow \infty$, then

$$\lim_{n\uparrow\infty} \left(\mathbf{I}+\mathbf{Q}+\mathbf{Q}^2+\cdots+\mathbf{Q}^{n-1}\right) = (\mathbf{I}-\mathbf{Q})^{-1} \tag{4.4.12}$$

Therefore,

$$\lim_{n\uparrow\infty} \mathbf{F}^{(n)} = (\mathbf{I}-\mathbf{Q})^{-1}\mathbf{R} \tag{4.4.13}$$

is a general formula for the matrix of absorption probabilities.

At this point a question that naturally arises is under what conditions may we conclude that $\mathbf{Q}^n \longrightarrow \mathbf{0}$ as $n \uparrow \infty$?. To answer this question, it will be helpful to consider a matrix norm defined as follows. Let $\mathbf{A} = (\mathbf{a}_{ij})$ be any finite square or rectangular matrix with real or complex elements. Then, define a matrix norm by

$$\| \mathbf{A} \| = \max_{\mathbf{i}} \sum_j \mid a_{ij} \mid . \tag{4.4.14}$$

Let $\mathbf{B}$ be any square or rectangular matrix such that the product $\mathbf{AB}$ is well defined. Then, it can be shown that

$$\| \mathbf{AB} \| \leq \| \mathbf{A} \| \times \| \mathbf{B} \| . \tag{4.4.15}$$

However, because

$$\sum_{i \in \mathfrak{S}} p_{ij} = 1 \tag{4.4.16}$$

for all $i \in \mathfrak{S}$ in the Wright-Fisher model under consideration, it follows that

$$\| \mathbf{Q} \| = \max_{i \in \mathfrak{S}_2} \sum_{j \in \mathfrak{S}_2} | p_{ij} |= \delta < 1. \tag{4.4.17}$$

Therefore,

$$\| \mathbf{Q}^n \| \leq \| \mathbf{Q} \|^n = \delta^n \tag{4.4.18}$$

so that

$$\lim_{n \uparrow \infty} \mathbf{Q}^n = \mathbf{0} \tag{4.4.19}$$

for the Wright-Fisher model under consideration, where $\mathbf{0}$ is a $r_2 \times r_2$ matrix of zeros. Moreover, the rate of convergence is geometric.

The result may be used to prove that the set of absorbing states will be entered eventually with probability one, given that the process starts in any transient state $i \in \mathfrak{S}_2$. For any integer $s \geq 1$, let $\mathbf{1}_s$ denote a $s \times 1$ vector of ones. For any $n \geq 1$, the *n-th* power of the transition matrix has the partitioned form

$$\mathbf{P}^n = \left(p_{ij}^{(n)} \right) = \begin{bmatrix} \mathbf{I}_{r_1} & \mathbf{0} \\ \mathbf{F}^{(n)} & \mathbf{Q}^n \end{bmatrix} . \tag{4.4.20}$$

It can be shown by induction that for all $n \geq 1$ that

$$\sum_{j \in \mathfrak{S}} p_{ij}^{(n)} = 1 \tag{4.4.21}$$

for all $i \in \mathfrak{S}_2$. Therefore, it follows that

$$\mathbf{F}^{(n)} \mathbf{1}_{r_1} + \mathbf{Q}^n \mathbf{1}_{r_2} = \mathbf{1}_{r_2} \tag{4.4.22}$$

for all $n \geq 1$. So if $\mathbf{Q}^n \to \mathbf{0}$ as $n \uparrow \infty$, then

$$\lim_{n \uparrow \infty} \mathbf{F}^{(n)} \mathbf{1}_{r_1} = \mathbf{1}_{r_2}. \tag{4.4.23}$$

Conversely, if this limit holds, then $\mathbf{Q}^n \to \mathbf{0}$ as $n \uparrow \infty$ so that the condition, $\mathbf{Q}^n \to \mathbf{0}$ as $n \uparrow \infty$, is necessary and sufficient for the process to enter the set of absorbing states eventually with probability one, given some initial transient state $i \in \mathfrak{S}_2$. It should be noted that this argument is valid only for the case the state space $\mathfrak{S}$ has finitely many members.

Let

$$\mathbf{F} = (f_{ij}) = \lim_{n\uparrow\infty} \mathbf{F}^{(n)}. \tag{4.4.24}$$

In the present age computers, which makes it possible to evaluate large matrices numerically, it is of interest to consider approaches for the numerical computation of the matrix **F**. As a first step, the recursive computation of the sequence of matrices

$$\left\{\mathbf{F}^{(n)} \mid n = 1, 2, \ldots\right\} \tag{4.4.25}$$

will be considered. It is easy to see that

$$\mathbf{F}^{(1)} = \mathbf{R} \tag{4.4.26}$$

and

$$\mathbf{F}^{(2)} = \mathbf{R} + \mathbf{Q}\mathbf{F}^{(1)}. \tag{4.4.27}$$

Furthermore, it can be seen that

$$\mathbf{F}^{(n)} = \mathbf{R} + \mathbf{Q}\mathbf{F}^{(n-1)} \tag{4.4.28}$$

is valid for $n \geq 2$. Whenever it is feasible to compute a large number of terms in the sequence (4.4.28) the resulting array of numbers would be of interest, because they may be interpreted as terms in a probability distribution function.

By letting $n \uparrow \infty$ in (4.4.28) it can be seen that the matrix **F** satisfies the set of linear equations

$$\mathbf{F} = \mathbf{R} + \mathbf{Q}\mathbf{F}. \tag{4.4.29}$$

Thus, the matrix **F** may be calculated by numerically solving the set of linear equations

$$(\mathbf{I} - \mathbf{Q})\,\mathbf{F} = \mathbf{R}. \tag{4.4.30}$$

Such sets of equations can often be solved efficiently by using a good computer implementation of the what is sometimes called the Gaussian elimination algorithm, which may be available in such software packages as MATLAB. If one uses this method, the inverse $(\mathbf{I} - \mathbf{Q})^{-1}$ may not be computed in the procedure. But, the value of **F** will be numerically equivalent to that in formula (4.4.30) expressed in terms of this inverse.

It is interesting to note that we now have enough information to describe the distribution of the random variable X in (4.3.15), which was the limit of a martingale sequence. The range of X is the set $\mathfrak{S}_1 = \{0, 2N\}$ of absorbing states and its conditional distribution, given that $X_0 = i \in \mathfrak{S}_2$, is as follows:

$$\begin{aligned} P\left[X = 0 \mid X_0 = i\right] &= f_{i0} \\ P\left[X = 2N \mid X_0 = i\right] &= f_{i,2N}. \end{aligned} \tag{4.4.31}$$

Given the distribution of the random variable X in (4.4.31), it is also very informative to exploit another martingale property. From (4.3.16), it follows that

$$E\left[X \mid X_0 = i\right] = 0 f_{i0} + 2N f_{i,2N} = i \tag{4.4.32}$$

so that

$$f_{i,2N} = \frac{i}{2N} \tag{4.4.33}$$

and

$$f_{i0} = 1 - \frac{i}{2N} \tag{4.4.34}$$

for all $i \in \mathfrak{S}_2$. In particular, if $i = N$, then

$$f_{i0} = f_{i,2N} = \frac{1}{2}. \tag{4.4.35}$$

In a terminology that is often used in genetics, the conditional probability in (4.4.33) would be interpreted as the probability an allele eventually becomes fixed in a population is equal to its initial frequency.

The simplicity of the fixation probabilities in the Wright-Fisher model is very interesting, but, they are not especially informative, because variation among realizations of the process is not reflected in these simple formulas. In subsequent sections, formulas for conditional expectations and variances will be derived within a framework of general finite Markov chains with absorbing states, which, when coupled with computer experiments, will provide insights into variation among realizations of the process.

There is also another of interpretation of the inverse of the matrix $(\mathbf{I} - \mathbf{Q})$ that is of interest. For any $j \in \mathfrak{S}_2$ define a sequence of indicator functions

$$\left(K_j^{(n)} \mid n = 0, 1, 2, \ldots\right) \tag{4.4.36}$$

as follows. If $X_n = j$, then let $K_j^{(n)} = 1$ and if $X_n \neq j$, let $K_j^{(n)} = 0$. Then, the random variable

$$T_j = \sum_{n=0}^{\infty} K_j^{(n)} \tag{4.4.37}$$

is the number of times, generations, the process is in transient state j prior the absorption in some absorbing state. Observe that for every pair of transient states (i, j)

$$\begin{aligned} E\left[K_j^{(n)} \mid X_0 = i\right] &= 1 \times P\left[X_n = j \mid X_0 = i\right] + 0 \times P\left[X_n \neq j \mid X_0 = i\right] \\ &= P\left[X_n = j \mid X_0 = i\right] = p_{ij}^{(n)}. \end{aligned} \tag{4.4.38}$$

Therefore, the conditional expectation of the number of times, generations, the process is in transient state j. Given the initial state $i \in \mathfrak{S}_2$, prior to absorption is

$$m_{ij} = E\left[T_j \mid X_0 = i\right] = \sum_{n=0}^{\infty} p_{ij}^{(n)}, \tag{4.4.39}$$

where, by definition, $p_{ij}^{(0)} = \delta_{ij}$, the Kronecker delta. In matrix form, this equation becomes

$$\mathbf{M} = (m_{ij}) = \sum_{n=0}^{\infty} \mathbf{Q}^n = (\mathbf{I} - \mathbf{Q})^{-1}. \tag{4.4.40}$$

Thus, whenever it is feasible to compute the matrix inverse on the right, the element m_{ij} may interpreted as the above conditional expectation $E\left[T_j \mid X_0 = i\right]$. Kemeny and Snell (1976) have also derived a formula for the matrix conditional variances $var\left[T_j \mid X_0 = i\right]$, for $i, j \in \mathfrak{S}_2$, but this formula will not be derived here.

4.5 Distributions of First Entrance Times Into an Absorbing State and Their Expectations and Variances

For every transient state $i \in \mathfrak{S}_2$ and absorbing state $j \in \mathfrak{S}_1$, let $g_{ij}^{(n)}$ denote the conditional probability that the process enters the absorbing state j at step or generation $n \geq 1$, given that the process started in transient state i. In symbols,

$$g_{ij}^{(n)} = P\left[X_n = j \mid X_\nu \neq j \text{ for } \nu = 1, 2, \ldots, n-1, X_0 = i\right]. \tag{4.5.1}$$

Define a $r_2 \times r_1$ matrix $\mathbf{G}^{(n)}$ by

$$\mathbf{G}^{(n)} = \left(g_{ij}^{(n)} \mid i \in \mathfrak{S}_2, j \in \mathfrak{S}_1\right). \tag{4.5.2}$$

Then, the relationship between these matrices and the matrix $\mathbf{F}^{(n)}$ of absorption probabilities is

$$\mathbf{F}^{(n)} = \sum_{\nu=1}^{n} \mathbf{G}^{(\nu)}. \tag{4.5.3}$$

It is also easy to see that for $n \geq 1$

$$\mathbf{G}^{(n)} = \mathbf{Q}^{n-1}\mathbf{R}. \tag{4.5.4}$$

In this formula, by definition, $\mathbf{Q}^0 = \mathbf{I}_{r_2}$, an identity matrix of order r_2. The rationale underlying this formula is the following. If the process entered an absorbing state for the first time in generation n, then in the preceding $n-1$ generations, the process remained in the set of transient states, and on the *n-th* step it jumped to an absorbing state and thereafter remained in that state. In a sense, this formula is a generalization of the scalar geometric distribution to the matrix case.

Let the random variable W_j denote the step or generation at which the process is absorbed into state $j \in \mathfrak{S}_1$. Then, given that $X_0 = i \in \mathfrak{S}_2$, the conditional expectation of this random variable is

$$\psi_{ij} = E\left[W_j \mid X_0 = i\right] = \sum_{n=1}^{\infty} n g_{ij}^{(n)}. \tag{4.5.5}$$

In matrix notation this formula becomes

$$\mathbf{\Psi} = (\psi_{ij}) = \sum_{n=1}^{\infty} n\mathbf{Q}^{n-1}\mathbf{R} = \left(\sum_{n=1}^{\infty} n\mathbf{Q}^{n-1}\right)\mathbf{R}. \tag{4.5.6}$$

Formally, the infinite series

$$\sum_{n=1}^{\infty} n\mathbf{Q}^{n-1} \tag{4.5.7}$$

reminds one of the scalar series

$$(1-x)^{-2} = \sum_{n=1}^{\infty} nx^{n-1}, \tag{4.5.8}$$

which is valid for $\mid x \mid < 1$. The validity of this series may be seen by differentiating the geometric series

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + \cdots \tag{4.5.9}$$

with respect to x. This formula suggests that equation

$$\sum_{n=1}^{\infty} n\mathbf{Q}^{n-1} = (\mathbf{I} - \mathbf{Q})^{-2} = \mathbf{M}^2 \tag{4.5.10}$$

is valid, where the matrix $\mathbf{M} = (\boldsymbol{I} - \boldsymbol{Q})^{-1}$ of conditional expectations was defined in the previous section. A proof of the validity of this formula may be constructed by formally considering the product

$$\left(\sum_{n=0}^{\infty} \mathbf{Q}^n\right) \times \left(\sum_{n=0}^{\infty} \mathbf{Q}^n\right) \tag{4.5.11}$$

of infinite matrix series but the details will be omitted.

From these results, it can be seen that the matrix $\boldsymbol{\Psi}$ of conditional expectations has the form

$$(E\left[W_j \mid X_0 = i\right]) = \boldsymbol{\Psi} = \mathbf{M}^2\mathbf{R}, \tag{4.5.12}$$

where $i \in \mathfrak{S}_2$ and $j \in \mathfrak{S}_1$. When r_2 is of moderate size, the matrix $\boldsymbol{\Psi}$ may be computed with relative ease on many computer platforms that will support software packages such as MATLAB and others. One may not be interested in all of the elements of the matrix $\boldsymbol{\Psi}$, but may be content with expected waiting times, expressed in generations, to entrance into some absorbing state. Let the random variable U denote the waiting time to entrance into some absorbing state. Then, the conditional probability density function of this random variable is

$$P\left[U = n \mid X_0 = i\right] = \sum_{j \in \mathfrak{S}_1} g_{ij}^{(n)}, \tag{4.5.13}$$

given that $X_0 = i$ for $n = 1, 2, \ldots$. Moreover, from this formula it can be seen that

$$\phi_i = E\left[U \mid X_0 = i\right] = \sum_{j \in \mathfrak{S}_1} \psi_{ij}. \tag{4.5.14}$$

Let $\boldsymbol{\phi}$ be a $r_2 \times 1$ vector with these elements and let $\mathbf{1}_{r_1}$ be a $r_1 \times 1$ vector of ones. Then, the vector $\boldsymbol{\phi}$ has the succinct vector-matrix form

$$\boldsymbol{\phi} = \boldsymbol{\Psi}\mathbf{1}_{r_1} = \mathbf{M}^2\mathbf{R}\mathbf{1}_{r_1}. \tag{4.5.15}$$

Another set of conditional expectations that are of interest are conditional expectations of the waiting time to absorption into absorbing state $j \in \mathfrak{S}_1$, given an initial transient state $i \in \mathfrak{S}_2$ and the event of final absorption into state j. Let A_j denote the event of final absorption into state j. Then, given the initial transient state i, the conditional probability of this event is

$$P\left[A_j \mid X_0 = i\right] = f_{ij}. \tag{4.5.16}$$

Therefore, the conditional expectation of the waiting time to absorption into state j, given the initial state i and the event A_j, is

$$E\left[Wj \mid X_0 = i, A_j\right] = \frac{1}{f_{ij}} \sum_{n=1}^{\infty} n g_{ij}^{(n)} = \frac{\psi_{ij}}{f_{ij}}. \tag{4.5.17}$$

In this equation we tacitly assume that $f_{ij} > 0$ for all pairs (i, j). When the computer programming is done is such array manipulating languages as APL, the computation of a matrix of elements in (4.5.17) is straight forward. Let

$$recip\mathbf{F} = \left(\frac{1}{f_{ij}}\right) \tag{4.5.18}$$

denote the reciprocal of the matrix $\mathbf{F} = (f_{ij})$. Then, the matrix of conditional expectations has the representation

$$\left(E\left[W_j \mid X_0 = i, A_j\right]\right) = \mathbf{\Psi} \times recip\mathbf{F} = (\mathbf{M}^2\mathbf{R}) \times recip\mathbf{F}, \tag{4.5.19}$$

where the multiplication of matrices on the right occurs element by element. Observe that in a computer implementation, the matrix $\mathbf{M}^2\mathbf{R}$ is computed first and then multiplied element by element with the matrix $recip\mathbf{F}$.

A derivation of a formula for the matrix of conditional variances

$$\left(var\left[W_j \mid X_0 = i, A_j\right]\right) \tag{4.5.20}$$

of the waiting times to absorption into the set of absorbing states would also be of interest as a measure of variation among realizations of a process. As a first step in this derivation, observe that

$$E\left[W_j^2 \mid X_0 = i, A_j\right] = E\left[W_j(W_j - 1) \mid X_0 = i, A_j\right] + E\left[W_j \mid X_0 = i, A_j\right]. \tag{4.5.21}$$

By definition,

$$E\left[W_j(W_j - 1) \mid X_0 = i, A_j\right] = \frac{1}{f_{ij}} \sum_{n=2}^{\infty} n(n-1) g_{ij}^{(n)}. \tag{4.5.22}$$

In matrix notation, this equation leads to the $r_2 \times r_1$ matrix

$$\left(E\left[W_j(W_j - 1) \mid X_0 = i, A_j\right]\right) = \left(\sum_{n=2}^{\infty} n(n-1)\mathbf{Q}^{n-1}\right) \mathbf{R} \times recip\mathbf{F}. \tag{4.5.23}$$

Therefore, a need to compute the matrix series

$$\sum_{n=2}^{\infty} n(n-1)\mathbf{Q}^{n-1} \tag{4.5.24}$$

in a finite number of operations arises. Fortunately, an approach to doing these calculations may be derived by appealing to properties of a scalar geometric series. To gain some insight into a procedure for summing the series in (6.24), consider the scalar series

$$(1-x)^{-2} = \sum_{n=1}^{\infty} nx^{n-1}, \tag{4.5.25}$$

which is valid of $\mid x \mid < 1$, see (6.8). A formal differentiation of the series in (6.25) leads to the expression

$$2\,(1-x)^{-3} = \sum_{n=2}^{\infty} n(n-1)x^{n-2}, \tag{4.5.26}$$

which is valid for $\mid x \mid < 1$.

This scalar series suggests that the matrix equation

$$2\mathbf{Q}\left(\mathbf{I}_{r_2} - \mathbf{Q}\right)^{-3} = \sum_{n=2}^{\infty} n(n-1)\mathbf{Q}^{n-1} = \mathbf{Q}\sum_{n=2}^{\infty} n(n-1)\mathbf{Q}^{n-2} \tag{4.5.27}$$

is valid whenever $\parallel \mathbf{Q} \parallel < 1$. That this indeed is the case may be proved formally but the details will be omitted. In summary, in order to compute the matrix series in (4.5.23), the matrix expression

$$2\mathbf{Q}\mathbf{M}^3 \tag{4.5.28}$$

needs to be evaluated numerically.

To derive a formula the matrix of conditional variances in (4.5.20), it will be helpful to start with the matrix equation

$$\begin{aligned}(var\,[W_j]) &= \left(E\left[\left[W_j^2\right]\right] - \left(E\left[\left[W_j\right]\right]\right)^2\right) \\ &= \left(E\left[W_j(W_j-1)\right] + E\left[W_j\right]\right) - \left(EW_j\right)^2 \end{aligned} \tag{4.5.29}$$

In order to simplify the notation, conditioning of the event $[X_0 = i, A_j]$ has been omitted. But, the equation would still be valid if conditioning this event had been explicitly included. To express this equation in a succinct matrix form, it will be useful to define another matrix operation.

For any matrix $\mathbf{A} = (a_{ij})$, the matrix of squares (a_{ij}^2) will be denoted by $\mathbf{A}_{sq}$. Given this definition, the matrix of conditional variances in (4.5.29) has the symbolic form

$$(var\,[W_j \mid X_0 = i, A_j]) = \left(2\mathbf{Q}\mathbf{M}^3\mathbf{R} + \mathbf{M}^2\mathbf{R}\right)\times recip\mathbf{F} - \left((\mathbf{M}^2\mathbf{R})\times recip\mathbf{F}\right)_{sq}. \tag{4.5.30}$$

At first sight, the formula in (4.5.30) calls for squaring and cubing the matrix $\mathbf{M}$, but, because this matrix can be rather large in the computer implementation of this formula, it is advantageous to search for ways to minimize the number of computations. For example for the Wright-Fisher model, when population size is $N = 500$, then $2N = 1000$ so that the matrix $\mathbf{M}$ is 999×999.

To this end, note that the matrix $\mathbf{F}$ of absorption probabilities is the product $\mathbf{F} = \mathbf{MR}$ and that this matrix is $r_2 \times r_1$. In many absorbing Markov chain models, $r_1 < r_2$. In the Wright-Fisher model, for example, $r_1 = 2$ and $r_2 = 2N - 1$ so that, in general, r_1 is small in comparison with r_2. This observation suggests that to compute $\mathbf{M}^2\mathbf{R}$ by the formula

$$\mathbf{F}_2 = \mathbf{MF} = \mathbf{M}^2\mathbf{R} \tag{4.5.31}$$

would be a computationally efficient procedure to follow, because it would entail fewer row to column multiplications and sums than in computing $\mathbf{M}^2\mathbf{R}$ in a two stage procedure by first computing $\mathbf{M}^2$. Similarly, the formula

$$\mathbf{F}_3 = \mathbf{MF}_2 = \mathbf{M}^3\mathbf{R} \tag{4.5.32}$$

would be an efficient way to compute $\mathbf{M}^3\mathbf{R}$. When expressed in terms of these equations, formula (6.30) has the more compact form

$$(var\,[W_j \mid X_0 = i, A_j]) = (2\mathbf{QF}_3 + \mathbf{F}_2) \times recip\mathbf{F} - (\mathbf{F}_2 \times recip\mathbf{F})_{sq}\,. \tag{4.5.33}$$

Observe that this expression is a $r_2 \times r_1$ matrix because $i \in \mathfrak{S}_2$ and $j \in \mathfrak{S}_1$.

The last formula to set down in this section is that for the vector of variances of the times of entrances into some absorbing state. Let $(var\,[U \mid X_0 = i])$ denote a $r_2 \times 1$ vector of these conditional variances. Then, this vector has the succinct representation

$$(var\,[U \mid X_0 = i]) = (2\mathbf{QF}_3 + \mathbf{F}_2)\mathbf{1}_{r_1} - (\mathbf{F}_2\mathbf{1}_{r_1})_{sq}\,. \tag{4.5.34}$$

When N is large, these conditional expectations and variances can be large so that the waiting time to absorption in some absorbing state can be very large, and, moreover, they may be a great deal of variation among the realizations of the process. This observation leads to the consideration of what has been called conditional quasi-stationary distributions on the set of transient states, given that the process remains in this set for the "long run".

There is a rather large literature on Wright-Fisher models. Chapter 3 of the interesting and informative book by Ewens (2004) may be consulted for additional details, where, unlike the presentation in this chapter, more attention is given to the eigenvalues and eigenvectors of transition matrices. But, as demonstrated by the ideas presented above, insights into Wright-Fisher models may be gained with only minimal mention of eigenvalues and eigenvectors of a matrix when powerful desk top computers are at one's disposal as was recognized by Kemeny and Snell (1976) at least 25 years ago.

When the population size N is large, then doing matrix calculations can become problematic. In this connection, applications of diffusion theory may lead to interesting results. For an extensive account of this theory, it is recommended that chapters 4 and 5 of Ewens (2004) be consulted and studied. An example of a useful approximation of the expected waiting time to absorption in some absorbing state is the formula

$$E\left[U \mid X_0 = i\right] \approx -4\left(p_i \ln p_i + (1-p_i)\ln(1-p_i)\right), \tag{4.5.35}$$

where

$$p_i = \frac{i}{2N}. \tag{4.5.36}$$

In the APL software, implementing some of the formulas in these notes, it is possible to judge the performance of this approximation formula by comparing it with the expectation on the left in (4.5.14).

4.6 Quasi-Stationary Distribution on the Set of Transient States

In the Wright-Fisher model under consideration, let E_n denote the event that the process is in the set of transient states at step or generation n. Then, the conditional probability of this event when the process starts in transient state i is

$$P\left[E_n \mid X_0 = i\right] = \sum_{j\in\mathfrak{S}_2} P\left[X_n = j \mid X_0 = i\right] = \sum_{j\in\mathfrak{S}_2} p_{ij}^{(n)}, \tag{4.6.1}$$

where $p_{ij}^{(n)}$ is an element of the matrix $\mathbf{Q}^n$. It follows, therefore, that

$$P\left[X_n = j \mid E_n, X_0 = i\right] = \frac{p_{ij}^{(n)}}{\sum_{\nu\in\mathfrak{S}_2} p_{i\nu}^{(n)}}, \tag{4.6.2}$$

for all pairs of transient states i, j. If there exists numbers π_{ij} such that

$$\lim_{n\uparrow\infty} P\left[X_n = i \mid E_n, X_0 = i\right] = \pi_{ij} \tag{4.6.3}$$

for all pairs i, j, then vector

$$(\pi_{ij} \mid j\in \mathfrak{S}_2) \tag{4.6.4}$$

is known as a quasi-stationary distribution, given that the initial transient state was i. By definition, the $r_2 \times r_2$ matrix

$$\mathbf{\Pi} \;=\; (\pi_{ij} \mid i \in \mathfrak{S}_2, j\in \mathfrak{S}_2) \tag{4.6.5}$$

will be called the set of quasi-stationary distributions on the set of transient states.

By appealing to the Perron-Frobenius theory of matrices with non-negative elements, it will be possible to not only show that the matrix $\mathbf{\Pi}$ exists but that it will also be possible to outline a procedure for calculating its elements. In an appendix, devoted to a review of matrix theory, Karlin and Taylor (1975), state and prove some theorems for matrices with non-negative elements. Let $\mathbf{A} = (a_{ij})$ denote a square matrix such that $a_{ij} \geq 0$ for all pairs i, j. The matrix is said to be positively regular if there is an integer $m > 0$ such that all elements of $\mathbf{A}^m$ are all positive. For the case of the Wright-Fisher model under consideration, the matrix $\mathbf{Q}$ in the partitioned form of the transition matrix has this property, because all its elements are positive, i.e., all elements satisfy the relation $p_{ij} > 0$ for all pairs (i, j).

When a matrix $\mathbf{A}$ is positively regular, according to the Perron-Frobenious theory, there exists an eigenvalue $\rho > 0$, of multiplicity one and such that if λ is any other eigenvalue, then $\mid \boldsymbol{\lambda} \mid < \boldsymbol{\rho}$. This eigenvalue is sometimes called the Perron-Frobenius root of $\mathbf{A}$. Suppose $\mathbf{A}$ is a $s \times s$ matrix for some integer $s > 0$. Then there is a $s \times 1$ eigenvector with elements x_i such that

$$\mathbf{Ax} = \boldsymbol{\rho}\mathbf{x}, \tag{4.6.6}$$

and it can be shown that $x_i > 0$ for all $i = 1, 2, \ldots, s$. The vector $\mathbf{x}$ is a right eigenvector corresponding to the eigenvalue value ρ. Similarly, let $\mathbf{y}$ be a $1 \times s$ left eigenvector with positive elements corresponding to the eigenvalue ρ such that

$$\mathbf{yA} = \rho\mathbf{y}. \tag{4.6.7}$$

It is possible to choose a constant such that the inner product satisfies the condition

$$\mathbf{yx} = \sum_{i=1}^{s} y_i x_i = 1. \tag{4.6.8}$$

A matrix that will play a key role in finding the matrix of quasi-stationary distributions is defined by

$$\mathbf{E} = \mathbf{xy} = (x_i y_j). \tag{4.6.9}$$

Observe that $\mathbf{E}$ is said to be a $s \times s$ idempotent matrix, because it has the property

$$\mathbf{E}^2 = \mathbf{E}. \tag{4.6.10}$$

To prove this note that

$$\mathbf{E}^2 = \mathbf{xyxy} = \mathbf{xy} = \mathbf{E}. \tag{4.6.11}$$

According to the appendix in Karlin and Taylor (1975), there is a number $a > 0$ such that $a < \rho$ and

$$\mathbf{A}^n = \rho^n \mathbf{E} + \mathbf{O}(a^n). \tag{4.6.12}$$

Therefore,

$$\lim_{n\uparrow\infty} \frac{1}{\rho^n} \mathbf{A}^n = \mathbf{E}. \tag{4.6.13}$$

In particular, consider the matrix $\mathbf{Q}$ and suppose ρ is the Perron-Frobenius root and that $\mathbf{x}$ and $\mathbf{y}$ are the right and left eigenvectors corresponding to ρ. Then,

$$\lim_{n\uparrow\infty} \frac{p_{ij}^{(n)}}{\sum_{\nu\in\mathfrak{S}_2} p_{i\nu}^{(n)}} = \lim_{n\uparrow\infty} \frac{p_{ij}^{(n)}/\rho^n}{\left(\sum_{\nu\in\mathfrak{S}_2} p_{i\nu}^{(n)}\right)/\rho^n} = \frac{x_i y_j}{\sum_{\nu\in\mathfrak{S}_2} x_i y_\nu} = \frac{y_j}{\sum_{\nu\in\mathfrak{S}_2} y_\nu} = \pi_{ij} \tag{4.6.14}$$

for pairs i, j of transient states. Observe that the ratio on the right does not depend on i and is entirely dependent on the left eigenvector $\mathbf{y}$.

Let $\mathbf{1}_{r_2}$ denote a $r_2 \times 1$ vector of ones and let $\boldsymbol{\pi}$ be $1 \times r_2$ vector with elements π_j defined by

$$\pi_j = \frac{y_j}{\sum_{\nu\in\mathfrak{S}_2} y_\nu} \tag{4.6.15}$$

for all $j \in \mathfrak{S}_2$. Then the matrix $\mathbf{\Pi}$ introduced above has the form

$$\mathbf{\Pi} = \mathbf{1}_{r_2}\boldsymbol{\pi}. \tag{4.6.16}$$

Observe that in this $r_2 \times r_2$ matrix, each row is the constant vector $\boldsymbol{\pi}$, which by definition, is the quasi-stationary distribution on the set of transient states $\mathfrak{S}_2$. It is important to observe that whatever the initial state $i \in \mathfrak{S}_2$, the limit

$$\lim_{n\uparrow\infty} P\left[X_n = j \mid E_n, X_0 = i\right] = \lim_{n\uparrow\infty} \frac{p_{ij}^{(n)}}{\sum_{\nu\in\mathfrak{S}_2} p_{i\nu}^{(n)}} = \pi_j \tag{4.6.17}$$

holds. The title, quasi-stationary distribution, is used to indicate that, although the process eventually reaches an absorbing state with probability one, when the waiting time to absorption is large, the process may tend to a stationary distribution in the sense of (4.6.17).

Given a matrix $\boldsymbol{Q}$, to find the quasi-stationary distribution, it would be necessary to find ρ, the Perron-Frobenius root of $\boldsymbol{Q}$. There are algorithms for calculating this root, but software implementing them may not be available. However, many software packages have program that will find all the eigenvalues of the matrix $\boldsymbol{Q}$, which usually perform well unless the matrix $\boldsymbol{Q}$ is too large. Fortunately, for the case of a Wright-Fisher process under consideration, all the eigenvalues of the matrix $\boldsymbol{Q}$ may be expressed in a symbolic form, see for example, Ewens (2004), where it is shown that all the eigenvalues of the transition matrix $\boldsymbol{P}$ have the symbolic form

$$\lambda_\nu = \frac{2N(2N-1)\cdots(2N-\nu+1)}{(2N)^\nu} \tag{4.6.18}$$

for $\nu = 1, 2, \ldots, 2N+1$. Any transition matrix has 1 as an eigenvalue, and, because there are two absorbing states, the root 1 has multiplicity 2. Therefore, all roots such that $\lambda_\nu \neq 1$ are the eigenvalues of the matrix $\boldsymbol{Q}$, which follows from the partitioned form of the transition matrix $\boldsymbol{P}$. From this remark, it can be seen from (4.6.18) that the eigenvalues of $\boldsymbol{Q}$ are

$$\lambda_\nu = \prod_{k=1}^{\nu-1}\left(1 - \frac{k}{2N}\right) \tag{4.6.19}$$

for $\nu \geq 2$. It is easy to see that $\lambda_2 > \lambda_3 > \cdots$ is a decreasing sequence. Therefore

$$\lambda_2 = 1 - \frac{1}{2N} = \rho \tag{4.6.20}$$

is the Perron-Frobenius root the matrix $\boldsymbol{Q}$. From this result, it follows that one may find the quasi-stationary distribution of for the Wright-Fisher model under consideration by using software that will compute the normalized left eigenvector of the matrix $\boldsymbol{Q}$ corresponding to the eigenvalue $\rho = 1 - 1/2N$.

By way of a simple example, suppose $N = 2$ so that $\boldsymbol{Q}$ is a 3×3 matrix. The eigenvalues of this matrix are $\rho = 1 - 1/4 = 3/4, \rho_1 = (3/4)\,(1-2/4) = 3/8$ and $\rho_2 = (3/8)(1-3/4) = 3/32$. By using the computation engine that is linked to the word processor, it can be shown that the quasi-stationary distribution in this case is

$$\boldsymbol{\pi} = \begin{pmatrix} \frac{8}{25} & \frac{9}{25} & \frac{8}{25} \end{pmatrix} = \begin{pmatrix} 0.32 \; 0.36 \; 0.32 \end{pmatrix}. \tag{4.6.21}$$

A quasi-stationary distribution may also exist for modification of a Wright-Fisher process in which mutation and selection has been incorporated into the model, as will be demonstrated in the next section. For such models, simple symbolic forms of the eigenvalues may not exist.

When such a process has at least one absorbing state, then it would be possible, in principle, to compute a quasi-stationary distribution if it were possible to compute the Perron-Frobenius root of the matrix $\boldsymbol{Q}$. One approach to finding this root is to compute all the eigenvalues of the matrix $\boldsymbol{Q}$ and then select the largest real eigenvalue. A limitation of this approach is that for large values of r_2, the number of transient states, the computation of all the eigenvalues may be problematic.

The algebra underlying matrix theory has a timeless quality, even though there have been recent advances in the theory motivated by the present computer age. A well written classical account of the theory of matrices with non-negative elements may be found in the book Gantmacher (1959).

An elegant account on the representation of square matrices as a sum of their principal idempotent matrices, which may be useful in studying the rate of convergence of the matrix $\boldsymbol{Q}$ to a zero matrix, can be found in the book Perlis (1952). An extensive account of symbolic forms of eigenvalues for transition matrices of Markov chains that arise in genetics may be found in the papers Cannings I and II (1974).

4.7 Incorporating Mutation and Selection Into Two Allele Wright-Fisher Processes

Mutation has been among the significant genetic forces driving evolution. In this section however, only one type of mutation will be considered. For the sake of simplicity, a closed system will be considered in that gene A_1 may mutate to gene A_2 and gene A_2 may mutate to gene A_1. It should be emphasized at the outset that a gene mutating into a new or different gene will not be accommodated in the formulation, although the incorporation of this feature into a formulation would be of considerable interest. A question as to whether a new mutant would survive in a population would also be of interest. To give an answer to this question, Fisher (1958) introduced a branching process into the formulation, and today, this process is widely known as the Galton-Watson process. There is a large literature on branching processes that will not be pursued here.

Let

$$\mathfrak{M} = \begin{bmatrix} \mu_{11} & \mu_{12} \\ \mu_{21} & \mu_{22} \end{bmatrix} \tag{4.7.1}$$

denote a matrix of mutation probabilities. In this matrix μ_{11} denotes the probability that gene A_1 does not mutate per generation and μ_{12} denotes the probability gene A_1 mutates to gene A_2 per generation. These probabilities satisfy the equation $\mu_{11} + \mu_{12} = 1$. The elements in the second row of the matrix are defined similarly. All the elements in the matrix $\mathfrak{M}$ belong to the interval $[0, 1] = [0 \leq x \leq 1]$.

Selection will be incorporated into the formulation by using the notion that selection acts at the gametic level so that a gene in a generation must be "selected", in a sense that will be explained subsequently, in order to enter the gene pool of the next generation. Let υ_1 and υ_2, respectively, denote the probabilities that genes A_1 and A_2 to enter the gene pool of the next generation. Both these numbers lie in the interval $[0 \leq x \leq 1]$.

By way of offering a biological example of this type of selection, it is of interest to consider the conception and birth histories of cohort of women, see Mode (1985) for citations of literature on this subject. It has been estimated that, among all conceptions experienced in a cohort of women, about 20% end in a spontaneous abortion. There may be many factors causing spontaneous abortions, and, among these some are thought to be genetic in nature. Suppose, for example, some mutant gene A_1 does affect the probability of spontaneous abortions. Then, if the fetus dies in utero, this gene would not become part of the gene pool of the next generation. In this sense, the gene would not be "selected". On the other hand, even though the fetus is a carrier of gene A_1, the pregnancy may go to full term and result in a live birth. Let υ_1 be the probability that this live birth survives and contributes to the gene pool of the next generation. Then, υ_1 is the probability that this gene is "selected".

Suppose that in generation n of a Wright-Fisher model, the state of the population is $X_n = i \in \mathfrak{S}$. Then, the probability that gene A_1 enters the gene pool of the next generation is

$$\eta_1(i) = \left(\left(\frac{i}{2N} \right) \mu_{11} + \left(1 - \frac{i}{2N} \right) \mu_{21} \right) \upsilon_1. \tag{4.7.2}$$

Similarly, the probability that gene A_2 enters the gene pool of the next generation is

$$\eta_2(i) = \left(\left(1 - \frac{i}{2N} \right) \mu_{22} + \left(\frac{i}{2N} \right) \mu_{12} \right) \upsilon_2. \tag{4.7.3}$$

Let $p_1(i)$ denote the conditional probability of finding gene A_1 in the gene pool of the population of the next generation. Then, by the law of total probability and Bayes' theorem, it follows that

$$p_1(i) = \frac{\eta_1(i)}{\eta_1(i) + \eta_2(i)}. \tag{4.7.4}$$

Similarly,

$$p_2(i) = \frac{\eta_2(i)}{\eta_1(i) + \eta_2(i)} \tag{4.7.5}$$

is the conditional probability of finding gene A_2 in the gene pool of the next generation. Observe that

$$p_1(i) + p_2(i) = 1 \tag{4.7.6}$$

for all $i \in \mathfrak{S}$.

Formula (4.7.5) was derived by defining the idea of selection in terms of probabilities and using well known concepts from probability theory. It should be pointed out, however, that other authors have formulated selection in terms of an idea of a selective advantage or selection coefficient which is denoted by a parameter $s \geq 0$. See, for example, the formula 2.4 on page 177 of the book by Karlin and Taylor (1981), where it is assumed that gene A_1 has a selective advantage over A_2. From an inspection of (4.7.5), it can be seen that the relationship of s to the probabilities of selection defined in this paper is

$$\frac{\upsilon_1}{\upsilon_2} = 1 + s \tag{4.7.7}$$

for $0 \leq \upsilon_1 \leq 1$ and $0 < \upsilon_2 \leq 1$. Observe that if $s = 0$, then $\upsilon_1 = \upsilon_2$ so that, by definition, there is no selection. Rather then working with the notion of selective advantage, from now on the idea of selection will be expressed in terms of probability concepts.

Given these formulas, the matrix of transition probability for the evolution of a Markov chain has the form

$$\mathbf{P} = (p_{ij}) = \left(\binom{2N}{j} (p_1(i))^j ((p_2(i))^{2N-j} \mid i \in \mathfrak{S}, \mathrm{j} \in \mathfrak{S} \right). \tag{4.7.8}$$

Given this parameterization of the model, there are three classes of cases that may be considered. If $\mu_{12} = \mu_{21} = 0$ and $\upsilon_1 = \upsilon_2$, then $p_1(i) = i/2N$ so that the model reduces a Wright-Fisher process with no mutation or selection. This case is also known as the neutral model. On the other hand, if $\upsilon_1 \neq \upsilon_2$, then selection may occur even if there is no mutation. This case gives rise to a Wright-Fisher process in which there is selection but no mutation. Another interesting class is that in which $\mu_{12} = 0$ but $\mu_{21} > 0$ so that there is mutation in only one direction; namely, A_2 mutates to A_1 but A_1 cannot mutate to A_2. For this class, if there is either no selection, $v_1 = v_2$, or selection, $v_1 \neq v_2$, then state $2N$ is the only absorbing state of the process and the transient states of the process are $i = 0, 1, 2, \ldots, 2N-1$. For this class of models, conditional expected waiting times to absorption as well as their standard deviations of these times may be computed using the techniques outlined in section (4.5).

A third class of cases arises when $\mu_{12} > 0$ and $\mu_{21} > 0$ so that there is mutation in both directions. In this class of models, either selection or no selection may be considered. For an inspection of formula (4.7.8) , it is easy to see that $p_{ij} > 0$ for all pairs (i, j) so that all states communicate and form a irreduceable class with no absorbing states. To simplify the notation let $r = 2N + 1$. From the general theory of this class Markov chains, see Karlin and Taylor (1981) or Chung (1960), it is known that there exists a stationary distribution $\boldsymbol{\pi}$ such that there is convergence

$$X_n \to \boldsymbol{\pi} \tag{4.7.9}$$

in distribution as $n \uparrow \infty$, where $\boldsymbol{\pi}$ is a $1 \times r$ vector of positive probabilities.

As in previous sections, let $\mathbf{1}_r$ denote a $r \times 1$ vector of ones. Then, it can be shown that

$$\lim_{n \uparrow \infty} \mathbf{P}^n = \mathbf{1}_r \boldsymbol{\pi}. \tag{4.7.10}$$

Thus, in the limit, as $n \uparrow \infty$, all rows in the matrix $\mathbf{P}^n$ tend to the vector $\boldsymbol{\pi}$. No attempt will be made to use a symbolic computation engine to investigate of properties of the irreduceable process under consideration. However, such an investigation may lead to new results. This limit result may also be deduced from the Perron-Frobenius theory of matrices with non-negative elements, but the proof of this statement will be left to the reader as an exercise.

Because each row of the matrix $\mathbf{P}$ satisfies the condition that it sums to one and all elements are positive, it follows that the Perron-Frobenius root of the matrix is $\rho = 1$. To see this, write this condition is vector matrix form

$$\mathbf{P}\mathbf{1}_r = \mathbf{1}_r. \tag{4.7.11}$$

From this equation, it can be seen that not only $\rho = 1$ is an eigenvalue of $\mathbf{P}$ but also that $\mathbf{1}_r$ is an eigenvector corresponding to this eigenvalue. If there exists another eigenvalue λ of this matrix such that $\mid \lambda \mid > 1$, then the above limit would not exist, which is contrary to theory. Hence, $\mid \lambda \mid < 1$.

Let $\mathbf{y}$ denote a $1 \times r_2$ left eigenvector corresponding to $\rho = 1$. Then,

$$\mathbf{y} = \mathbf{yP}. \tag{4.7.12}$$

According to the Perron-Frobenious theory, there is a solution to this homogeneous equation such that all the elements of $\mathbf{y} = (y_i \mid i = 1, 2, \ldots, r)$ are positive, *i.e.*, $y_i > 0$ for all i. Then, the normalized vector $\boldsymbol{\pi}$ with elements

$$\pi_i = \frac{y_i}{\sum_{\nu=1}^{r_2} y_\nu} \tag{4.7.13}$$

is the stationary distribution of the process. If there is a computer program for finding the left vector $\mathbf{y}$ corresponding to the root $\rho = 1$, then the stationary distribution could, in principle, be calculated. Various software packages and programming languages have such capabilities, including, for example, MATLAB and APL 2000. In a subsequent section, examples of these computations will be given.

There are also other interesting quantities that may be calculated from the stationary distribution of the process. Let η_{ii} denote the expected recurrence time of state $i \in \mathfrak{S} = (i \mid i = 0, 1, 2, \ldots, 2N)$, and let T_{ii} denote a random variable such that if the process enters state i at some time, then T_{ii} is the number of generations the process reenters state i. Then, it is known that $E[T_{ii}] = \nu_{ii}$ is given by

$$\nu_{ii} = \frac{1}{\pi_i} \tag{4.7.14}$$

for all states $i \in \mathfrak{S}$. For proofs of this result, the books on stochastic processes cited above may be consulted. In computer experiments with this model, it will, therefore, be of interest to compute the quantities in (4.7.14) and display them in a graph as a function of the state of the process.

4.8 Genotypic Selection with no Mutation and Random Mating

Genotypic selection is, by definition, the case in which it is assumed that selection acts at the level of the genotype. This is a case that has not been considered in foregoing sections of this chapter. When there are two alleles at a locus, then there are three possible genotypes; namely the homozygotes A_1A_1 and A_2A_2 and the heterozygote A_1A_2. Throughout this section, $\mu_{12} = \mu_{21} = 0$ so that mutation is not considered. If p_n is the frequency (probability) of gene A_1 in some population in generation $n = 0, 1, 2, \ldots,$

then, under the assumption of random mating, the frequencies of the three genotypes in the population are given in the second column of the following table

A_1A_1	p_n^2
A_1A_2	$2p_nq_n$
A_2A_2	q_n^2

(4.8.1)

where $q_n = 1 - p_n$. It is assumed in what follows that the inequality $0 < p_n < 1$ holds for all n.

Let v_{11} denote the probability that genotype A_1A_1 is selected to contribute offspring to generation $n + 1$, and the probabilities v_{12} and v_{22} be defined similarly for genotypes A_1A_2 and A_2A_2, respectively If $v_{11} = v_{12} = v_{22} = 1$, then by definition there is no selection. Because $v's$ are defined as probabilities, it follows that of selection probabilities lie in the closed interval $[0, 1]$. If for any genotype the probability of selection is zero, then a gene or genes carried by this genotype is called a lethal. For example, if A_1 is a lethal gene but there must be two copies of the gene for to express lethality, then $v_{11} = 0$ but $v_{12} > 0$ and $v_{22} > 0$. It will also be assumed in what follows that all selection probabilities are constant from generation to generation.

The probabilities that each of the three genotypes are selected to contribute gametes to the next generation are given by the second column of the following table

A_1A_1	$p_n^2v_{11}$
A_1A_2	$2p_nq_nv_{12}$
A_2A_2	$q_n^2v_{22}$

. (4.8.2)

From this column, it follows that

$$W_n = p_n^2v_{11} + 2p_nq_nv_{12} + q_n^2v_{22} \tag{4.8.3}$$

is the total probability that some genotype contributes a gamete to the next generation. Some authors call the function W_n the mean fitness of the population in generation n.

Let p_{n+1} denote the conditional probability that a gamete in generation $n+1$ carries the gene A_1. Then, by applying the Bayes' rule, it follows that

$$p_{n+1} = \frac{p_n^2v_{11} + p_nq_nv_{12}}{p_n^2v_{11} + 2p_nq_nv_{12} + q_n^2v_{22}} \tag{4.8.4}$$

for $n = 0, 1, 2, \ldots$. Given some initial frequency p_0 of gene A_1 and assigned values of the selection probabilities, this equation may be used as a deterministic model to calculate the frequency p_n of gene A_1 in the gene pool of the population for all generations $n \geq 1$. For the case that the heterozygote is more "fit"; i.e., $v_{12} > v_{11}$ and $v_{12} > v_{22}$, to contribute genes to the gene pool of the next generation, it is known that the limit

$$\lim_{n \uparrow \infty} p_n = \widehat{p} \tag{4.8.5}$$

exists and is a function of the selection probabilities under some conditions that will be stated subsequently. In the genetic literature, this is called a balanced polymorphism, because selection pressure maintains both genes A_1 and A_2 in the population and the frequency of gene A_1 in this equilibrium population is $\widehat{p}$. Some authors refer to this case as that of an infinite population, because, unlike Wright-Fisher model, the size of the population is not taken into account in this deterministic formulation.

In a Wright-Fisher formulation, however, the size N of a diploid populating is assumed to be constant from generation to generation and the formulation of a model of selection at the genotypic level takes the following form. Let

$$p_n(i) = \frac{i}{2N} \tag{4.8.6}$$

denote the frequency of allele A_1 in generation n, when the process is in transient state $i = 1, 2, \ldots, 2N - 1$. Then the probability $p_{n+1}(i)$ in the binomial density

$$p_{ij} = \binom{2N}{j} (p_{n+1}(i))^j \left(q_{n+1}(i)\right)^{2N-j}, \tag{4.8.7}$$

for $j = 0, 1, 2, \ldots, 2N$, is determined by above recursive formula (4.8.4) for every $n \geq 1$.

Just as in the model for gametic selection described in a previous section, there are two absorbing states, $(0, 2N)$, and $2N - 1$ transient states $(i \mid i = 1, 2, \ldots, 2N - 1)$. Therefore, with probability one, either allele A_1 or A_2 becomes fixed in the population so in the limit as $n \uparrow \infty$ the population does not reach an equilibrium or balanced polymorphism such that both alleles are maintained in the population by selection. This phenomena clearly differentiates the stochastic and deterministic formulations under consideration. In computer experiments with a Wright-Fisher formulation, the computer output for any experiment would be the same as that described in previous sections.

It is of some interest to derive a formula for the equilibrium probability $\widehat{p}$ for the deterministic formulation. The value of this probability is a solution of the equation

$$p = \frac{p^2 v_{11} + pqv_{12}}{p^2 v_{11} + 2pqv_{12} + q^2 v_{22}}, \tag{4.8.8}$$

where $q = 1 - p$. Interestingly, the symbolic computation engine linked to this word processor gives the following symbolic solutions to this equation:

$$\left\{ \begin{array}{cc} \left\{0, 1, \frac{-v_{12}+v_{22}}{v_{11}-2v_{12}+v_{22}}\right\} \text{ if} & v_{11} - 2v_{12} + v_{22} \neq 0 \\ \mathbb{C} & \text{if } -v_{12} + v_{22} = 0 \wedge v_{11} - 2v_{12} + v_{22} = 0 \\ \{0, 1\} & \text{if } -v_{12} + v_{22} \neq 0 \wedge v_{11} - 2v_{12} + v_{22} = 0 \end{array} \right\}. \tag{4.8.9}$$

Evidently, the symbol $\mathbb{C}$ means the solutions of the equation are not well defined when the indicated conditions are satisfied. In a balanced polymorphism, the frequency of allele A_1 would be

$$\widehat{p} = \frac{v_{12} - v_{22}}{2v_{12} - v_{11} - v_{22}}. \tag{4.8.10}$$

if the condition

$$2v_{12} - v_{11} - v_{22} > v_{12} - v_{22} > 0 \tag{4.8.11}$$

is satisfied. Observe that this condition ensures that $0 < \widehat{p} < 1$.Interestingly, for the stochastic model under consideration, it is not necessary that condition (4.8.11) be satisfied.

By way of an example in using this formula, suppose $v_{11} = 0.9, v_{12} = 1$ and $v_{22} = 0.89$, then

$$\widehat{p} = \frac{1 - 0.89}{2 - 0.9 - 0.89} = 0.523\,809\,523\,8. \tag{4.8.12}$$

It is of interest to note that the deterministic formulation is a three-parameter model with v_{11}, v_{12} and v_{22} as the parameters. On the other hand, the Wright-Fisher stochastic formulation is a four-parameter model in that this formulation has an additional parameter N, the population size. For this model, it would be of interest to compute the quasi-stationary distribution of the process, because, in a sense, this distribution characterizes a balanced polymorphism for the case of a Wright-Fisher process accommodating genotypic selection but no mutation.

4.9 A Computer Experiment with the Wright-Fisher Neutral Model

As was indicated in a foregoing section, when mutation and selection are taken into account in a Wright-Fisher process with two autosomal alleles, then the neutral case within this structure may be derived by letting $\mu_{12} = \mu_{21} = 0$ and $v_1 = v_2 = 1$ so that there is no mutation or selection. As was shown in section 4.4, for the neutral case there are two absorbing states, 0 and $2N$, where N is population size. The set $\mathfrak{S}_2$ of transient states for the neutral case is, therefore, $\mathfrak{S}_2 = (i \mid i = 1, 2, \ldots, 2N - 1)$.

In section 4.5, formulas for the expected waiting times and their standard for entrance into some absorbing state were derived. A diffusion approximation to expected waiting time to absorption in some absorbing state was also displayed in (4.5.35). A computer program was written to compute many of the matrices and expectations described in section 4.5. More precisely, the output of this program is a $(2N - 1) \times 6$ such that the first column is the set of transient states and the second and third columns are the absorption probabilities, given any transient state $i \in \mathfrak{S}_2$, see (4.4.13).

The fourth column is the conditional expectations given in (4.5.14), and the fifth column was computed by using approximation formula (4.5.35) for every transient state $i \in \mathfrak{S}_2$. Finally, the sixth column is the conditional standard deviations computed from formula (4.5.34). In this section, the results of a computer experiment will be reported for population size $N = 500$. It is interesting to note that these calculations involved inverting the 999×999 matrix $\boldsymbol{I} - \boldsymbol{Q}$. The quasi-stationary distribution was also computed for the case $N = 500$. In this calculation, knowing that the Perron-Frobenius root of $\boldsymbol{Q}$ was $1 - (2N)^{-1}$ was very helpful.

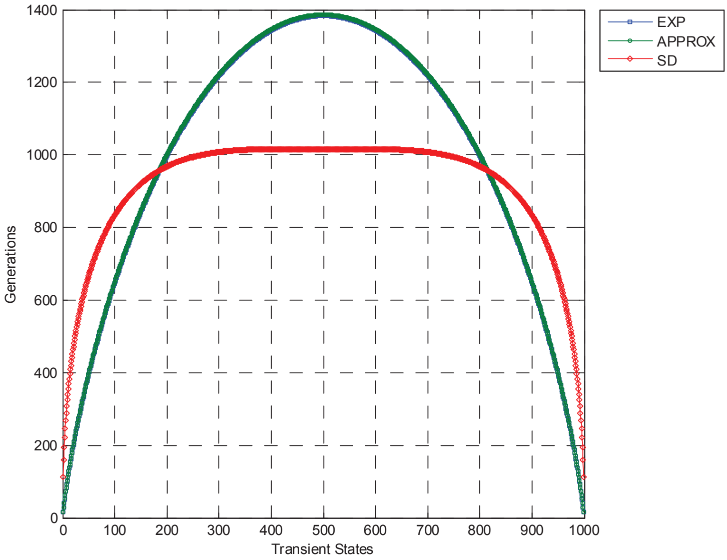

Figure 4.9.1 Conditional Expectations and Standard Deviations of Waiting Times in Generations to Fixation in an Absorbing State.

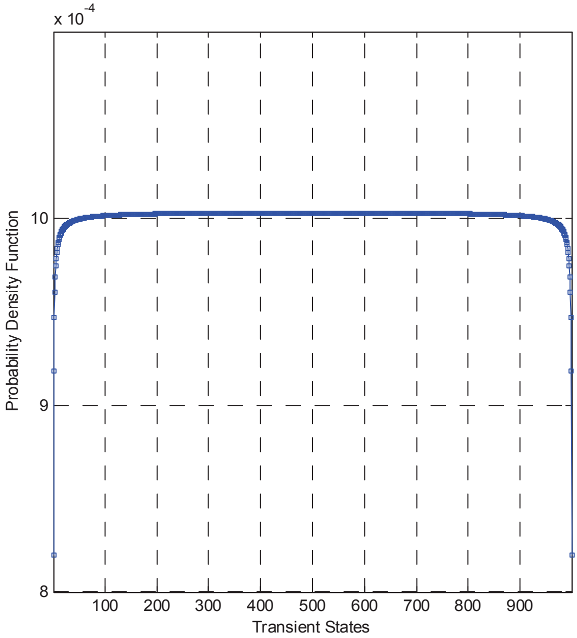

Figure 4.9.2 Probability Density Function of the Quasi-Stationary Distribution of a Neutral Wright-Fisher Process with Population Size 500.

Presented in Figure 4.9.1 are the graphs of the expectations (EXP) and the standard deviations (SD) conditioned on each transient state, $1, 2, \ldots, 999$, and the diffusion approximation (APPROX) to the expectations. As may be seen from these graphs, the expectation and the approximation are, for all practical purposes, the same.

It is interesting to note that for transient states 200 to 800, the conditional expectation is greater than the standard deviation, but, outside this range, SD is greater than EXP. As can be seen from the graph, when initial population size is 500, the conditional expectation of the waiting time to fixation is about 1,400 generations.

Figure 4.9.2 contains the graph of the density function of the quasi-stationary distribution of a neutral Wright-Fisher process. It is interesting to note, that this graphs of the probability density function suggests that quasi-stationary distribution on the transient states is nearly uniform except for points near the extreme transient states 1 and 999, which reflect the tendency for fixation in one of the absorbing states 0 and $2N = 1,000$.

4.10 A Computer Experiment with Wright-Fisher Selection Model

Another case of interest arises when it is assumed that there is no mutation in a Wright-Fisher model with two autosomal alleles but one allele, say A_1 may have a selective advantage. For such cases, $\mu_{12} = \mu_{21} = 0$ and $\upsilon_1 > \upsilon_2$. The parameter values chosen to do the calculations in this section were $N = 75, \upsilon_1 = 0.99$ and $\upsilon_2 = 0.9$. For the class of Wright-Fisher processes under consideration no general formulas of the eigenvalues of the $\boldsymbol{Q}$-matrix seem to be known, when mutation and selection are taken into account. Therefore, if one wishes to compute the quasi-stationary distribution of the process, it is necessary to compute all the eigenvalues of $\boldsymbol{Q}$ and then find the maximum of their absolute values. In exploratory experiments, it was found that when $N > 75$, convergence to all the eigenvalues of the matrix may not occur. However, for the parameter values listed above, convergence did occur so that it was possible to find the Perron-Frobenius root of the matrix $\boldsymbol{Q}$. For the case, $N = 75$, the dimensions of the matrix $\boldsymbol{Q}$ are 149×149. Because the diffusion approximation to the conditional expected waiting times to absorption in some state do not seem to be known for cases with selection and mutation that are accommodated in the model, it was possible to compute these expectations for only those cases in which the matrix formulas in section 4.5 were applicable.

Figure 4.10.1 contains the graphs of the absorption probabilities of fixation in states 0 or $2N$ for 75 probabilities for absorbing states 0 and $2N$, conditioned on the initial transient state. Because the allele A_1 is assumed to have a selective advantage, the conditional probability of absorption in state $2N$ rises quickly to 1 as one moves from left to right on the set of transient states $(1, 2, \ldots, 149)$. Similarly, the conditional probability of fixation in absorbing 0, decreases quickly as one moves from left to right on the set of transient states.

Presented in Figure 4.10.2 are the graphs of the conditional expectations and standard deviations of the waiting times in generations to fixation in some absorbing state, conditioned on the initial transient state. As can be seen from these graphs, the graph of the expectation EXP is uniformly higher than that for standard deviation SD but they both increase as a function of the initial transient state. This decreasing nature of these graphs indicates that selection drives allele A_1 rather rapidly to fixation with decreasing variation in the times to fixation as a function of the initial transient state. As can be seen, for example, from the graphs, when

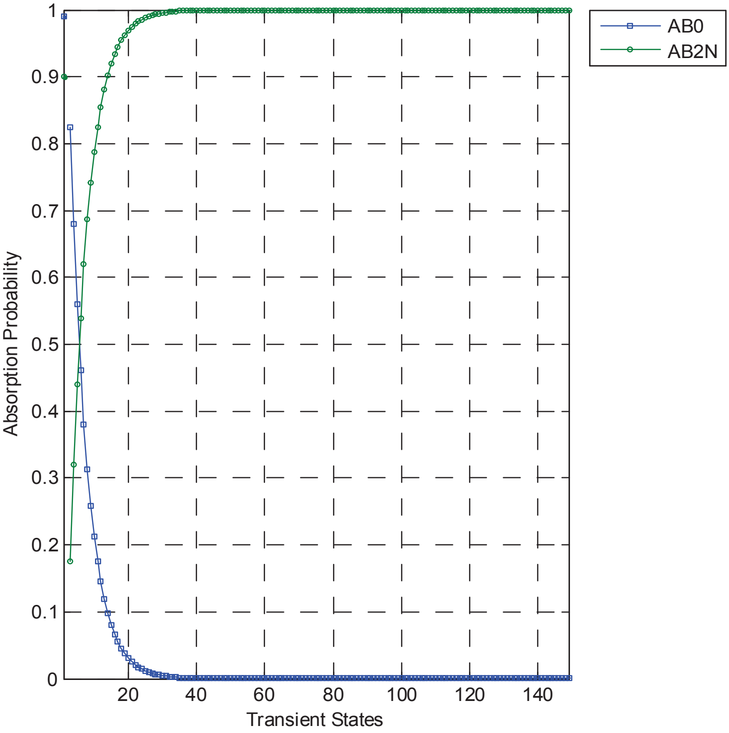

Figure 4.10.1 Probabilities of Fixation in State 0 or 150 as Functions of the Initial Transient State.

the initial transient state in near 149, then the expected waiting time to fixation is about 10 generations.

Figure 4.10.3 contains a graph of the distribution function of the quasi-stationary distribution on the set of transient states $(1, 2, \ldots, 149)$. As can be seen from this graph, this distribution is skewed to the right, indicating that, given that fixation of allele A_1 has not occurred, gene frequencies in the range 120/150 to 149/150 are most probable.

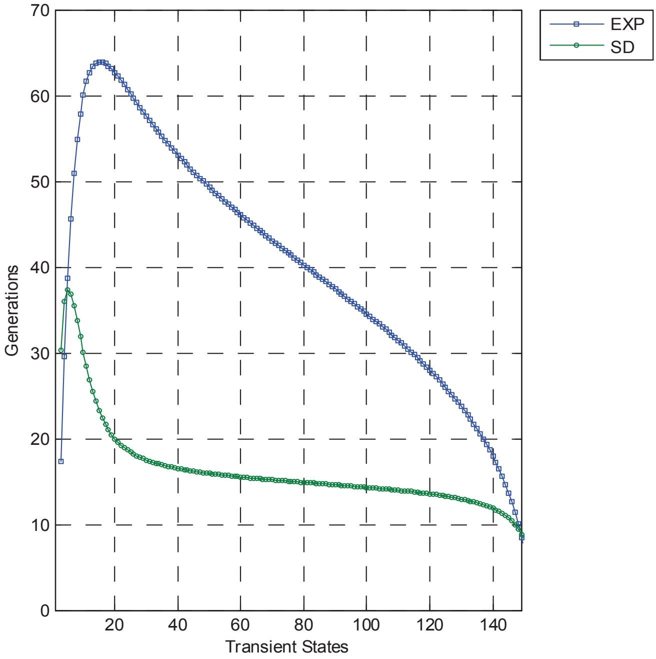

Figure 4.10.2 Expectations and Standard Deviations of Waiting Times to Fixation as Functions of the Initial Transient State.

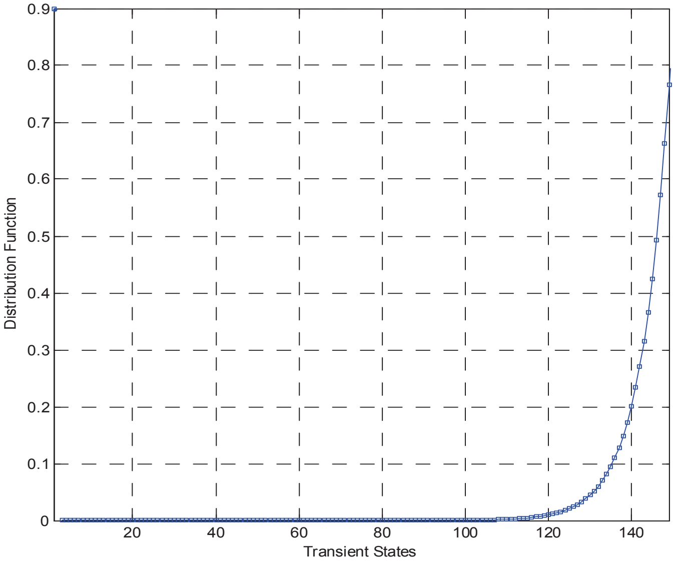

Figure 4.10.3 Quasi-Stationary Distribution of the Set of Transient States.

4.11 A Computer Experiment with Wright-Fisher Genotypic Selection Model

This section will be devoted to a computer experiment with the Wright-Fisher genotypic selection model with no mutation as described in section 4.8. The population size chosen for this experiment was $N = 75$. The reason for choosing this value was a desire to calculate the quasi-stationary distribution of the process. This calculation in turn required that all the eigenvalues of the matrix $\boldsymbol{Q}$ be computed so that its Perron-Frobenius root could be determined. The values chosen for the selection probabilities of the three genotype A_1A_1, A_2A_2 and A_1A_2 were $\upsilon_{11} = 0.9, \upsilon_{22} = 0.8$ and $\upsilon_{12} = 0.95$. Given these values for the parameters, it was possible to compute not only absorption probabilities and conditional waiting times to absorption with their standard deviations but also the quasi-stationary distribution.

Figure 4.11.1 contains the graphs of the probabilities of fixation in absorbing state 0 and 150 as functions of the initial transient state. Viewed

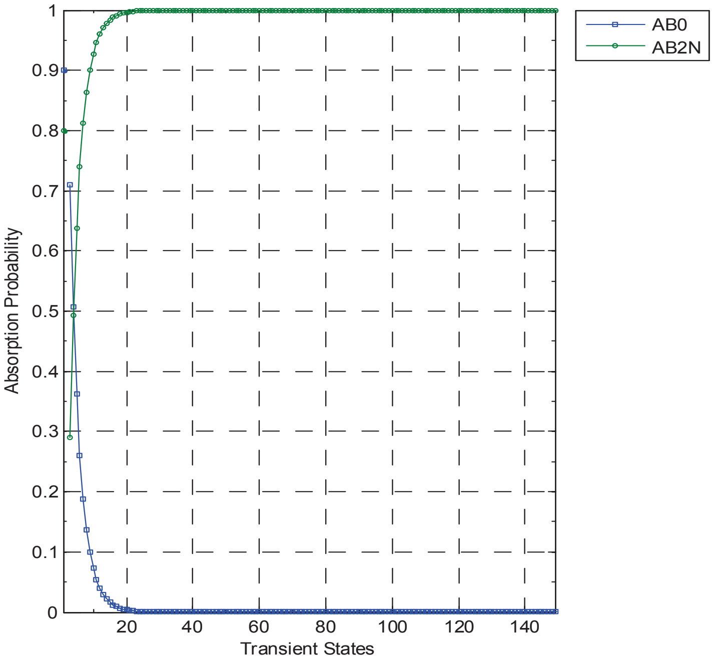

Figure 4.11.1 Probabilities of Fixation as Functions of the Initial Transient State.

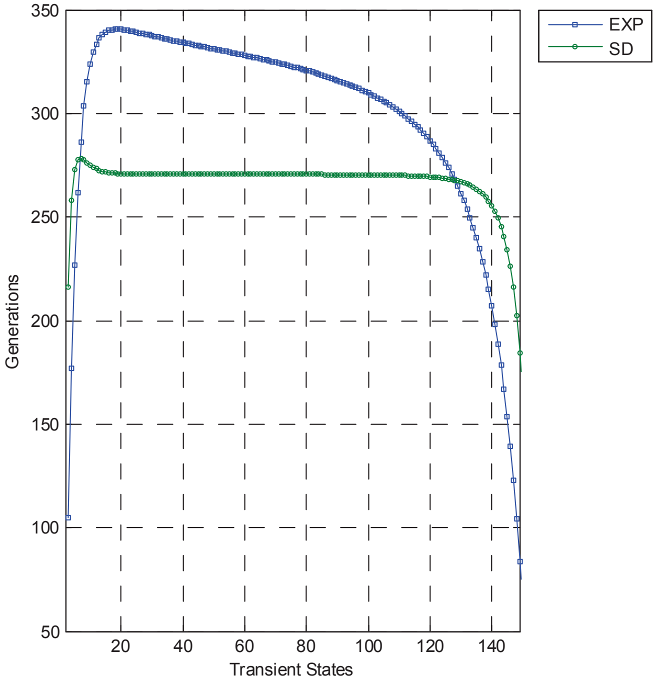

Figure 4.11.2 Expectations and Standard Deviations of Waiting Times to Fixation as Functions of the Initial State.

together, these graphs display quantitative aspects of allele A_1 having a selective advantage over A_2. In this connection, observe that the probability of fixation in state $2N = 150$ increases rapidly to 1 as a function of the initial transient state, which indicates that the frequency of allele A_1 may increase rather rapidly as the population evolves. However, the rate of increase in frequency depends on the initial transient state. At the same time, the probability of fixation in state 0, indicating that the population is homozygous for allele A_2, decreases rapidly to 0 as a function of the initial transient state.

Contained in Figure 4.11.2 are the graphs of the waiting times to fixation in an absorbing state as a function of the initial transient state. Like the graphs in Figure 4.10.2, those in this figure increase and then decrease as a function of the initial transient state. But, in this case, the decrease in these graphs is slower than in Figure 4.10.2, indicating the selection acts in such a way that both alleles are present in the population for a greater number of generations than those for the model studied in section 4.10.

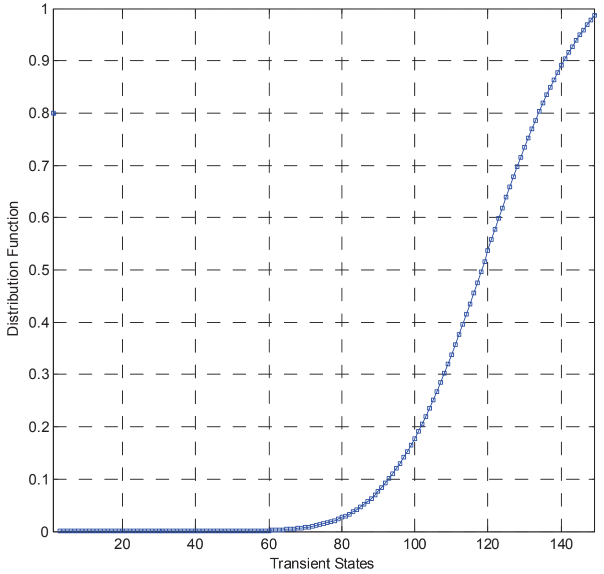

Figure 4.11.3 Quasi-Stationary Distribution of the Set of Transient States.

Figure 4.11.3 contains a graph of the distribution function of the quasi-stationary distribution on the set of transient states $(1, 2, \ldots, 149)$. Like the quasi-stationary distribution in Figure 4.10.3, the one in this figure is skewed to the right, indicating the selection is resulting in an increase in the frequency of allele A_1, prior to its fixation in the population. For this model, range of most probable frequencies of allele A_1 is wider than that for the model studied in section 4.10. For from figure 4.11.3, it can be seen that the most probable frequencies of allele A_1 are in the range 80/150 to 149/150.

4.12 A Computer Experiment with a Wright-Fisher Model Accommodating Selection and Mutation

As was indicated in a previous section, when there is mutation and selection in a Wright-Fisher model, then the process will converge eventually to a stationary distribution which may be computed by finding the normalized left eigenvector corresponding to the Perron-Frobenius root $\rho = 1$ of transition matrix of the process as set forth in section 4.7. In this section, the results of computer experiment for the parameter values

$N = 500, \mu_{12} = \mu_{21} = 0.00001$, $\upsilon_1 = 0.954$ and $\upsilon_2 = 0.953$. Observe that the values of the selection parameters differ only at the third decimal point so that allele A_1 has only a slight selective advantage over allele A_2.

Presented in Figure 4.12.1 is a graph of the expected recurrence times ν_{ii} in generations for all states for all the states $i \in \mathfrak{S} = (0, 1, 2, \ldots, 1000)$, the state space of the process. The expectation ν_{ii} is the expected number of generations for the process in statistical equilibrium to return to state i, given that in some generation the process was in state i. Recall, that if π_i for $i \in \mathfrak{S}$ is the stationary distribution of the process, then $\nu_{ii} = \pi_i^{-1}$ so that if ν_{ii} is large, then π_i is small.

From the graph in Figure 4.12.1, it can be seen that ν_{ii} as a function of $i \in \mathfrak{S}$, increases from state 0 to a maximum of about 5×10^4 generations near state 300 and then declines to a smaller values near state 1000. As an exercise, the reader may wish to visualize the graph of the stationary distribution by looking at the relationship $\pi_i = \nu_{ii}^{-1}$. From this relationship, one can see that the probabilities in the stationary distribution are largest near the states 0 and 1,000. Furthermore, the distribution is skewed to the right, which reflects the slight selective advantage of allele A_1

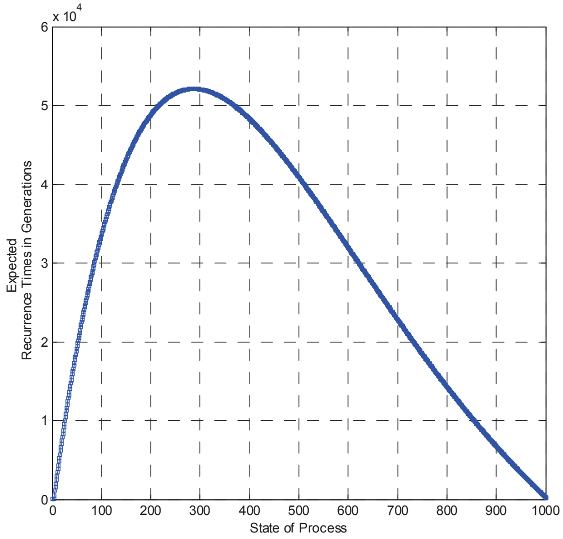

Figure 4.12.1 Expected Recurrence Times for a Process in Statistical Equilibrium

Bibliography

[1] Cannings, C. (1974) The Latent Roots of Certain Markov Chains Arising in Genetics: New Approach I. Haploid Models. **Advances in Applied Probability 6:**264–290.

[2] Cannings, C. (1974) The Latent Roots of Certain Markov Chains Arising in Genetics: A New Approach, II. Further Haploid Models. **Advances in Applied Probability 7:**260–282.

[3] Chung, K. L. (1960) **Markov Chains with Stationary Transition Probabilities.** Springer, Berlin, Heidelberg, New York.

[4] Ewens, W. J. (2004) **Mathematical Population Genetics I. Theoretical Introduction**. Springer, New York, London, Paris, Tokyo. (second edition)

[5] Fisher, R. A. (1958) **The Genetical Theory of Natural Selection**. Dover, New York. (second revised edition)

[6] Gantmacher, F. R. (1959) **The Theory of Matrices vol II**. Chelsea, New York.

[7] Karlin, S., Taylor, H. M. (1975) **A First Course in Stochastic Processes**. Academic Press, Boston, New York, London. (second edition)

[8] Karlin, S., Taylor, H. M. (1981) **A Second Course in Stochastic Processes**. Academic Press, Boston, New York, London.

[9] Kemeny, J. G., Snell, J.. L. (1976) **Finite Markov Chains**. Springer, New York, London.

[10] Mode, C. J. (1985) **Stochastic Processes in Demography and Their Computer Implementation**. Springer, Berlin, Heidelberg and New York.

[11] Perlis, S. (1952) **Theory of Matrices**. Addison-Wesley, Reading, Mass, USA.

Chapter 5

Multitype Gamete Sampling Processes, Generation of Random Numbers and Monte Carlo Simulation Methods

5.1 Introduction

In this chapter, various extensions of the Wright-Fisher processes accommodating two autosomal alleles discussed in chapter 4 will be formulated and investigated in illustrative computer simulation experiments. The models introduced in this chapter may be classified as multitype gamete sample processes, which include models with multiple alleles as in Mendelian genetics or other types of multitype gametes involving genes or regions of DNA on two or more chromosomes. Among the models to be described in this chapter is a preliminary formulation of a model for inherited autism, involving six types of gametes, and designed to accommodate mutations on any of the 22 pairs of autosomal chromosomes in the human genome. In addition to considering a straight-forward extension of the two-allele Wright-Fisher process studied in chapter 4 to cases of three or more multiple alleles at some locus, a class of models, derived from branching processes by conditioning on total population size, will also be formulated and studied in computer simulation experiments.

Even when cases of three types of gametes are considered the number of states in a state space $\mathfrak{S}$ of the Markov chain becomes so large that it is no longer feasible to implement the types of matrix calculations described in chapter 4 on present-day desk top computers. As will be demonstrated in this chapter however, samples of Monte Carlo realizations of such processes may be calculated and analyzed statistically with relative ease. A problem that arises in connection with the execution of Monte Carlo simulation experiments, entailing the computation of large numbers of random numbers, is that of documenting whether the random number generator used in the experiments passes a battery empirical tests for statistical randomness. In the spirit of scientific openness, the random number generator to be used throughout this book was documented in considerable detail is a section devoted to the mathematics and statistics underlying the theory of random number generators. Also included in this chapter is a section in which the statistical techniques used to construct informative summarizations of Monte Carlo simulation data are documented in considerable detail.

5.2 A Wright-Fisher Model with Multiple Types of Gametes – Mutation and Selection

In this section, it will be assumed that there are $k > 2$ types of gametes under consideration. These types of gametes may represent multiple Mendelian alleles at some autosomal locus or they may represent combinations of alleles at loci on either one or two or more autosomal chromosomes. In subsequent sections, examples of various types of gametes will be given. In this section, however, the types of gametes under consideration will be denoted by the symbols A_ν for $\nu = 1, 2, \ldots, k$. Just as in the case of two alleles at one locus discussed in the previous chapter, for the multiple gametes model under consideration, some among the k types of gametes may arise from new mutations. Let N denote the size of a diploid population, which will be assumed to be constant for all generations. Then, in any generation there will be $2N$ gametes and let the non-negative integer $i_\nu \geq 0$ denote the number of copies of gamete type A_ν for $\nu = 1, 2, \ldots, k$. Then, because the number of gametes in the population is, by assumption, constant from generation to generation, it follows that equation

$$i_1 + i_2 + \cdots + i_k = 2N \tag{5.2.1}$$

will be satisfied in every generation. For the case of multiple gametes, the state space of the Markov chain that characterizes a Wright-Fisher model in this case is the set $\mathfrak{S}$ of all solutions of equation (5.2.1). Symbolically, this set may be defined as

$$\mathfrak{S} = (\,(i_1, i_2, \ldots, i_k) \mid i_1 + i_2 + \cdots + i_k = 2N), \tag{5.2.2}$$

where $i_\nu = 0, 1, 2, \ldots, 2N$ for $\nu = 1, 2, \ldots, k$.

The set $\mathfrak{S}$ may contain a very large number M of elements, and by using a well-known combinatorial formula, the exact number of elements in this set may be computed as a function of k and N. As was shown in chapter 1, this number is given by the formula

$$M = \binom{2N + k - 1}{k}. \tag{5.2.3}$$

By way of an illustration, if $N = 500$ and $k = 3$, then

$$M = 167,\ 167,\ 000. \tag{5.2.4}$$

As illustrated by this example, the number of states in the set $\mathfrak{S}$ can be very large, which would preclude the computation of a $M \times M$ Markov transition matrix on most desktop computers. Subsequently, it will be shown that the problem of designing a Monte Carlo simulation procedure for the model under consideration is, in essence, straight-forward; consequently, it will be feasible to implement multiple gametes models on many desktop computers if k is not too large.

As a first step in formulating a multiple allele model with mutation and selection, let

$$\mathfrak{M} = (\mu_{\nu\nu'}) \tag{5.2.5}$$

denote a $k \times k$ matrix of mutation probabilities per generation. If $\nu \neq \nu'$, then $\mu_{\nu\nu'}$ is the probability gene A_ν mutates to gene $A_{\nu'}$, but, if $\nu = \nu'$, then $\mu_{\nu\nu}$ is the probability that gamete A_ν does not mutate. All the elements of this matrix belong to the interval $[0, 1]$ and for every $\nu = 1, 2, \ldots, k$

$$\sum_{\nu'=1}^{k} \mu_{\nu\nu'} = 1. \tag{5.2.6}$$

In some generation n, suppose the population is in state $\mathbf{i} = (i_1, i_2, \ldots, i_k) \in \mathfrak{S}$, and let

$$\mathbf{x}(\mathbf{i}) = \left(\frac{i_\nu}{2N} \mid \nu = 1, 2, \ldots, k \right) \tag*{(5.2,7)}$$

denote the corresponding $1 \times k$ vector of gamete frequencies. Given state $\mathbf{i}$, let $\eta_\nu(\mathbf{i})$ denote the probability of finding gamete A_ν in the potential gene pool of generation $n + 1$ after mutation has occurred, and let $\boldsymbol{\eta}(\mathbf{i}) = (\eta_1(\mathbf{i})), \eta_2(\mathbf{i}), \ldots, \eta_k(\mathbf{i}))$ denote a $1 \times k$ vector of these probabilities. Then,

$$\boldsymbol{\eta}(\mathbf{i}) = \mathbf{x}(\mathbf{i})\,\mathfrak{M}. \tag{5.2.8}$$

To accommodate selection in the model, let υ_ν denote the probability that gamete A_ν is selected to enter the gene pool of generation $n+1$. Then, given $\mathbf{i}$, $\eta_\nu(\mathbf{i})\,\upsilon_\nu$ is the probability gamete A_ν is in the gene pool of generation

$n+1$. Given the state $\mathbf{i}$ in generation n, let $p_\nu(\mathbf{i})$ denote the conditional probability that gene A_ν is in the gene pool of generation $n+1$. Then, by an application of the law of total probability and Bayes' theorem, it follows that

$$p_\nu(\mathbf{i}) = \frac{\eta_\nu(\mathbf{i})\,\upsilon_\nu}{\sum_{\nu=1}^{k}\eta_\nu(\mathbf{i})\,\upsilon_\nu}. \tag{5.2.9}$$

Let the vector random variable $\mathbf{X}_n$ denote the state of the population for every generation $n = 1, 2, \ldots$. Then, for every pair of states $(\mathbf{i}, \mathbf{j}) \in \mathfrak{S} \times \mathfrak{S}$, let

$$P[\mathbf{X}_{n+1} = \mathbf{j} \mid \mathbf{X}_n = \mathbf{i}] = p(\mathbf{i}, \mathbf{j}) \tag{5.2.10}$$

denote the stationary transition probabilities of a Markov chain for all $n = 0, 1, 2, \ldots$. For the case of a Wright-Fisher process with k types of gametes, it is assumed that the random sampling of gametes from generation to generation is characterized by multinomial distribution. More precisely, for every pair of states $(\mathbf{i}, \mathbf{j}) \in \mathfrak{S} \times \mathfrak{S}$, with the vector $\mathbf{j} = (j_1, j_2, \ldots, j_k)$,

$$p(\mathbf{i}, \mathbf{j}) = \frac{(2N)!}{\prod\limits_{\nu=1}^{k}(j_\nu)!}\prod_{\nu=1}^{k}(p_\nu(\mathbf{i}))^{j_\nu}. \tag{5.2.11}$$

Consequently, the problem of designing a Monte Carlo simulating procedure to compute realizations of a Wright-Fisher process with multiple alleles, reduces to computing realizations of samples from multinomial distributions. As was shown in chapter 1, the problem of computing realizations of a random vector with a multinomial distribution reduces to the problem of computing realizations from conditional binomial distributions, see sections 1.9 and 1.11.

The general formulation of a Wright-Fisher process with multiple gametes just described contains many special cases, but only a few will be mentioned here. If there is no mutation so that the mutation matrix has the form $\mathfrak{M} = \mathbf{I}_k$, a $k \times k$ identity matrix, and all selection probabilities are equal $\upsilon_1 = \upsilon_3 = \cdots = \upsilon_k$, then a so-called neutral Wright-Fisher process arises with k absorbing states. In particular, the set of absorbing states is the set of k vectors $((2N, 0, \ldots, 0), (0, 2N, 0, \ldots, 0), \ldots, (0, 0, \ldots, 0, 2N))$, indicating that only one of the k types of gametes becomes fixed eventually with probability one. For the neutral model, if $\boldsymbol{i} \in \mathfrak{S}$ is the state of the population in some generation n, then the evolution of the next generation is governed by a multinomial distribution with probability vector

$$\boldsymbol{p}(\boldsymbol{i}) = \left(\frac{i_\nu}{2N} \mid i_1 + i_2 + \cdots + i_k = 2N; \nu = 1, 2, \ldots, k\right) \tag{5.2.12}$$

and index $2N$. If $i_\nu/2N$ is the initial frequency of gamete A_ν in a population, then the probability that this allele is eventually fixed in the population is $i_\nu/2N$. To prove these statements, it would suffice to use a slight alteration of the proofs discussed in chapter 4 for a Wright-Fisher process with two autosomal alleles. As can be demonstrated in a Monte Carlo simulation experiment, if population size N is large, then in the neutral case the frequencies of the k alleles in a population may not change significantly over thousands of generations. Such persistence of gene frequencies over many generations is often referred to as genetic drift.

A second case of interest is that in which all elements of the mutation matrix $\mathfrak{M}$ are positive and all probabilities of selection are positive but may or may not be equal. Then, all M states in the state space $\mathfrak{S}$ of the Markov chain form a recurrent class so that, given any initial state $\mathbf{i} \in \mathfrak{S}$, the process will eventually converge in distribution to a stationary distribution, which may, in principle, be computed as the left eigenvector corresponding to $\rho = 1$, the Perron-Frobenius root of the transition with probabilities defined as in (5.2.11). But, in most cases of interest, the transition matrix of the Markov chain will be too large to carry out the calculations needed to compute the stationary distribution of the process. For those readers who are interested in coalescence theory and the use of a Dirichlet as a stationary distribution for a Wright-Fisher model with multiple alleles, the paper of Griffiths and Tavare (1994) may be consulted. This paper is also of interest from the standpoint of applications of the Monte Carlo methods, where they are used to solve large systems of linear equations. It seems likely that such methods to solve large systems of linear equations will in future be used more extensively.

Among the many cases of the multiple gametes model under consideration that may be entertained is a three gamete model with a mutation matrix of the form

$$\mathfrak{M} = \begin{pmatrix} \mu_{11} & \mu_{12} & \mu_{13} \\ 0 & \mu_{22} & \mu_{23} \\ 0 & 0 & 1 \end{pmatrix}. \tag{5.2.13}$$

This matrix is chosen such that either of the mutation $A_1 \to A_2$ or $A_1 \to A_3$ may occur with positive probability per generation, and, moreover, the mutation $A_2 \to A_3$ may occur with positive probability. "A" after a mutation to allele A_3 occurs there is no mutation to either allele A_1 or A_2. For this particular model, the state $(0, 0, 2N)$, indicating that gamete type A_3 has been fixed in a population, is the only absorbing state. The paper, Mode and Gallop (2008) may be consulted for the results of a Monte Carlo simulation experiment based on this model, where it was assumed gamete A_3 has a selective advantage over the other types of gametes.

Before proceeding to other examples, it should be mentioned that in this section mutation and selection have been formulated in terms of classical Mendelian genetics. That is, more detailed information of the concept of a gene has not been incorporated into the formulas. In this connection, before an attempt is made to incorporate more of the fine structure of human DNA into a formulation, it would be advisable to consult such treatises on Human Molecular Genetics such as that of Strachan and Read (2004), including such topics as mutation and DNA repair, chromosomal anomalies, the identification of human disease genes in terms of their structure at the molecular level and coding regions of DNA.

5.3 Examples of Multiple Alleles and Types of Gametes Involving Two Chromosomes

In genetics, the term multiple alleles is used to describe a situation in which a gene at some Mendelian locus has more that two alleles. In chapter 1, the A, B, AB and O blood types in humans were mentioned as an example of multiple alleles in connection with applications of the multinomial distribution in genetics. For this example, it has been determined that three alleles, denoted by I^A, I^B and i, govern the inheritance of blood type. Alleles I^A and I^B are both dominant to allele i but are also codominant, *i.e.*, individuals of genotype $I^A I^B$ are such that both the A and B antigens may be found in their blood. Further details on this system may be found in the recent text book, Snustad and Simmons (2006). Another example of multiple alleles described in this text book are those governing coat color in rabbits. By making extensive crosses among homozygotes, it has been determined that four alleles govern coat color in rabbits. The allele c^+ (wild-type) is dominant to the other three alleles, which are denoted by c (albino) , c^h (himalayan) and c^{ch} (chinchilla). Moreover, the chinchilla allele is partially dominant to himalayan and albino alleles and the himalayan allele is completely dominant to the albino allele. The relationships among these alleles is sometimes represented in the symbolic form $c^+ > c^{ch} > c^h > c$. Cases of multiple alleles form one class of the multiple types of gametes model with selection and mutation described in section 5.2. In this connection, it is thought by some investigators that the three alleles for coat color and markings are mutants of the wild type allele c^+.

Another class of the types of model discussed in section 5.2 arises when the case two loci on separate chromosomes are considered with two alleles per locus. Actually, more than two alleles at each locus could be entertained, but for the sake of simplicity only cases of two alleles at each locus will be considered. Let A_ν and B_ν for $\nu = 1, 2$ denote the two alleles at each locus. Then, in this illustrative model, there would be four types of gametes; namely, A_1B_1, A_1B_2, A_2B_1 and A_2B_2. To identify these four types with a single subscript, let $\kappa_1 = 11, \kappa_2 = 12, \kappa_3 = 21$ and $\kappa_4 = 22$. Then, for the sake of brevity, the four types of gametes will be denoted by the symbols κ_ν for $\nu = 1, 2, 3, 4$. To accommodate mutation in this example, one needs to specify the elements in the 4×4 mutation matrix $\mathfrak{M} = \left(\mu_{\kappa_v \kappa_{\nu'}}\right)$.

Since probabilities of mutation per generation may differ on the two chromosomes, it seems reasonable to consider mutation probabilities for each locus under consideration. Let τ_{12} denote the probability allele A_1 mutates to allele A_2 and let τ_{21} denote the probability of the mutation $A_2 \longrightarrow A_1$. Similarly, let the symbols ς_{12} and ς_{21} denote the probabilities of the mutations $B_1 \longrightarrow B_2$ and $B_2 \longrightarrow B_1$, respectively. Then, if it is assumed that mutations at the two loci are independent events, it follows that the probability $\mu_{\kappa_1\kappa_1} = (1 - \tau_{12})(1 - \varsigma_{12})$. By similar reasoning, the probability of the mutation $\kappa_1 \longrightarrow \kappa_2$ is $\mu_{\kappa_1\kappa_2} = (1 - \tau_{12})\varsigma_{12}$. The same kind of rationale may be used to specify all 16 elements in the 4×4 mutation matrix $\mathfrak{M}$. Thus, columns 1 and 2 of the matrix $\mathfrak{M}$ have the symbolic form

$$\begin{pmatrix} (1-\tau_{12})(1-\varsigma_{12}) & (1-\tau_{12})\varsigma_{12} \\ (1-\tau_{12})\varsigma_{21} & (1-\tau_{12})(1-\varsigma_{21}) \\ \tau_{21}(1-\varsigma_{12}) & \tau_{21}\varsigma_{12} \\ \tau_{21}\varsigma_{21} & \tau_{21}(1-\varsigma_{21}) \end{pmatrix}, \tag{5.3.1}$$

and columns 3 and 4 of $\mathfrak{M}$ have the symbolic form

$$\begin{pmatrix} \tau_{12}(1-\varsigma_{12}) & \tau_{12}\varsigma_{12} \\ \tau_{12}\varsigma_{21} & \tau_{12}(1-\varsigma_{21}) \\ (1-\tau_{21})(1-\varsigma_{12}) & (1-\tau_{21})\varsigma_{12} \\ (1-\tau_{21})\varsigma_{21} & (1-\tau_{21})(1-\varsigma_{21}) \end{pmatrix}. \tag{5.3.2}$$

The matrix $\mathfrak{M}$ must have the property that the sum of each row is 1. To see that this is indeed the case for the first row, observe that the sum of the first row in matrix (5.3.1) is $1 - \tau_{12}$, and the sum of the first row of matrix (5.3.2) is τ_{12}. Hence, $1 - \tau_{12} + \tau_{12} = 1$. By similar arguments, it can be shown that all the rows of the matrix $\mathfrak{M}$ sum to 1 as they should.

Selection may be incorporated into this model with 4 types of gametes by specifying four probabilities of selection, υ_ν for $\nu = 1, 2, 3, 4$ and proceeding as in section 5.2.

5.4 A Genetic Theory for Inherited Autism in Man

Autism Spectrum Disorder (ASD) in man is characterized by such behavioral traits as language impairments, social deficits and repetitive behaviors and may occur with only one or several children in a family. A family with only one child diagnosed with ASD is said to be a case of sporadic autism; while those of two or more children with ASD are said to be cases of familial or inherited autism. Over all, the incidence of ASD is about 1 in 150 births, according to information released by the Untied States Centers for Disease Control. Evidence for the presence of genetic factors that may be implicated in the disease has been provided by studies of identical or monozygotic twins, where it has been observed that if one twin is diagnosed with ASD, then with probability 0.7 or greater the other twin also has the disorder. There is also evidence for the involvement of genetic factors in dizygotic twins or fraternal twins, but, for such cases, if one twin is diagnosed with ASD, then the other twin is less likely to be diagnosed with the condition than with monozygotic twins, and, according to some estimates, the probability that both twins are autistic may be less than 0.1. A recent paper by Zhao *et al.* (2007) may be consulted for references and other information on the above statements concerning ASD.

According to these authors, it is likely that autism involves many genes, because linkage studies, directed towards attempts to implicate segments of DNA on different chromosomes, have been found to have no major effects. Furthermore, it has been found that there is only minor allele sharing over the entire genome when segments DNA of sibs, brothers and sisters, with ASD are compared. There is also evidence from cytogenetic studies that the number of copies of genes present in an individual may also be implicated in ASD. Another factor that may be implicated in autism is the presence of spontaneous mutations in the gametes of parents that they pass on to their children who may be diagnosed with ASD.

In this connection, the authors entertained a model in which spontaneous denovo mutations, which do not remain in the gene pool of a population for long periods of time because of selection, were invoked to explain cases of sporadic and inherited autism in families. A basic assumption underlying this model is that, phenotypically, autism in males, caused by one or more spontaneous mutation events at one or more loci throughout the genome, is inherited as a Mendelian dominant with high penetrance but with lower penetrance in females. Males with severe cases of autism often fail to attract sexual partners so that they do not have any offspring and thus do not pass on these mutations to their children. Thus, in such cases, denovo mutations are eliminated from the population as part of the process of natural selection. Such mutational events were hypothesized to account for cases of sporadic autism in families. On the other hand, because there is a lower degree of penetrance in females who may carry denovo mutations, such females may attract sexual partners and may, therefore, pass on the mutations to their offspring, who may be in turn be diagnosed with ASD. According to the authors, such cases may account for familial or inherited autism. In the next section, a model will be constructed to study the plausibility of the model just outlined within gamete sampling framework.

5.5 An Evolutionary Genetic Model of Inherited Autism

As a first step in developing a model to study the implications of the structure suggested in section 5.4 within a Wright-Fisher evolutionary gamete sampling process, let A_1 denote a type of gamete such that at least one mutation is present in the genome that may lead to autism, and let A_2 denote a gamete that is free of mutations that may cause autism. It will be assumed that A_2 may mutate to type A_1 but, by assumption, there is no mutation from A_1 to A_2. Let μ_{21} denote the probability per generation that there is a mutation from A_2 to A_1 and let μ_{22} denote the probability there is no mutation. As a first step in developing a formula for μ_{21}, consider only those mutations that may occur on one or more of the 22 pairs autosomal chromosomes that are present in the human genome. Let μ denote the probability per generation that some mutation occurs on some autosomal chromosome that may lead to autism. Then, assuming mutations on different chromosomes occur independently, it follows that $\mu_{22} = (1-\mu)^{44}$ is the probability that no mutations occur among the 22 autosomal chromosomes. Hence

$$\mu_{21} = 1 - (1-\mu)^{44} \tag{5.5.1}$$

is the probability that at least one mutation occurs on some of the 44 chromosomes under consideration. Actually, the mutation probability μ may differ in males and females, but for the sake of simplicity, such differences will be ignored in the preliminary formulation that will be described in this section. Another factor that will be omitted is that the parameter μ may depend on the age of the parents.

To accommodate the idea of penetrance in the model, let π_m denote the probability that a male who carries the mutation in A_1develops autism. If it is assumed that A_1 acts as a Mendelian dominant, then an offspring with either genotype A_1A_1 or A_1A_2 is at risk of developing autism. In what follows, the phenotype that includes both these genotypes will be denoted by A_1-. By definition, among males of phenotype A_1-, the fraction π_m will develop autism and the fraction $1-\pi_m$ will not develop autism. The symbol π_f will denote penetrance of autism in females who carry the mutation A_1 and has the same definition as that for males. In general, $\pi_m > \pi_f$ so that there is greater penetrance for autism in males than in females for individuals who carry the mutation A_1.

From the perspective of gametic frequencies in a diploid population with N individuals, let p_f denote the probability that at birth a child is female, and let $p_m = 1 - p_f$ denote the probability that a child is male at birth. Given that a child is female, let q_{fA_1} denote the conditional frequency of gene A_1 among those females with symptoms of autism, and let $q_{fA_{1-}}$ denote the frequency of gene A_1 in those females who do not have symptoms of autism. Finally, let q_{fA_2} denote the frequency of gene A_2 among females in the population. By definition, $q_{fA_1} + q_{fA_{1-}} + q_{fA_2} = 1$. The frequencies $q_{mA_1}, q_{mA_{1-}}$ and q_{mA_2} are defined similarly for males. With respect to a Wright-Fisher process with multiple types of gametes, let the 1×6 vector

$$\boldsymbol{p}_n = \left(p_f q_{fA_1}, p_f q_{fA_{1-}}, p_f q_{fA_2}, p_m q_{mA_1}, p_m q_{mA_{1-}}, p_m q_{mA_2}\right) \qquad (5.5.2)$$

denote the symbolic frequencies of the 6 types of gametes in a population in some generation n.

In what follows, it will be assumed that among those females and males in generation n with symptoms of autism, natural selection acts in such a way that they are not able to attract mates, and, therefore, will not contribute offspring to a subsequent generation. Under this assumption, the frequencies of alleles A_1 and A_2 among those females who may attract mates in generation n are

$$\boldsymbol{r}_{nf} = \frac{1}{q_{fA_{1-}} + q_{fA_2}} \left(q_{fA_{1-}}, q_{fA_2}\right) = \left(r_{nfA_1}, r_{nfA_2}\right). \qquad (5.5.3)$$

Similarly, among those males who may attract mates in generation n, the frequencies of alleles A_1 and A_2 are

$$\boldsymbol{r}_{nm} = \frac{1}{q_{mA_{1-}} + q_{mA_2}} \left(q_{mA_{1-}}, q_{mA_2} \right) = \left(r_{nmA_1}, r_{nmA_2} \right). \tag{5.5.4}$$

As was done in previous chapter 3, when representing any genotype symbolically, the allele on the left will be interpreted as that allele contributed by the female and that on the right allele contributed the male. Under the assumption of random mating, frequencies of the four types of genotypes A_1A_1, A_1A_2, A_2A_1 and A_2A_2 in generation n among females and males that will contribute offspring to the next generation. However, the genotypes A_1A_2 and A_2A_1 express the same phenotype so that the frequencies of these two genotypes will be combined as shown in the table 5.5.1.

Table 5.5.1 Genotypes and Their Frequencies Under Random Mating

Genotypes	**Frequencies**
$A_1A_1 = g_1$	$r_{nfA_1}r_{nmA_1} = s_{n1}$
$A_1A_2 = g_2$	$2r_{nfA_1}r_{nmA_2} = s_{n2}$
$A_2A_2 = g_3$	$r_{nfA_2}r_{nmA_2} = s_{n3}$

Observe that in order to simplify the notation in what follows, from now on the three genotypes and their frequencies will be represented by one letter with a subscript as indicated in table 5.5.1

Let $\boldsymbol{s} = (s_{n1}, s_{n2}, s_{n3})$ denote a 1×3 vector whose elements are the frequencies of the three genotypes in table (5.5.1). By assumption, the frequency of these genotypes within the female and male populations in generation n is the same. Therefore, under the assumption of random mating, the frequency of a mating of type (g_i, g_j) among the matings in generation n who produce offspring making up generation $n+1$ is given by $s_i s_j$. Consequently, the distribution of the 9 types of matings among females and males in generation n that will contribute offspring to the next generation, is given the elements in the 3×3 matrix

$$\boldsymbol{s}^T \boldsymbol{s}. \tag{5.5.5}$$

The 9 types of matings with frequencies in (5.5.5) however, may be lumped into six types of matings and their frequencies as listed in the table 5.5.2, where the genotypes of females and males in a mating are not distinguished.

Table 5.5.2 Types and Frequencies of Matings

Mating	**Frequency**
$g_1 \otimes g_1$	s_{n1}^2
$g_1 \otimes g_2$	$2s_{n1}s_{n2}$
$g_1 \otimes g_3$	$2s_{n1}s_{n3}$
$g_2 \otimes g_2$	s_{n2}^2
$g_2 \otimes g_3$	$2s_{n2}s_{n3}$
$g_3 \otimes g_3$	s_{n3}^2

The conditional distributions of the three genotypes among the children in generation $n+1$ from each type of mating type in generation n are presented in the table 5.5.3.

Table 5.5.3 Genotypic Distributions for Each Type of Mating

Mating	g_1	g_2	g_3
$g_1 \otimes g_1$	1	0	0
$g_1 \otimes g_2$	$\frac{1}{2}(1+\mu_{21})$	$\frac{1}{2}\mu_{22}$	0
$g_1 \otimes g_3$	μ_{21}	μ_{22}	0
$g_2 \otimes g_2$	$\frac{1}{4}(1+\mu_{21})^2$	$\frac{1}{2}(1+\mu_{21})\mu_{22}$	$\frac{1}{4}\mu_{22}^2$
$g_2 \otimes g_3$	$\frac{1}{2}(1+\mu_{21})\mu_{21}$	$\frac{1}{2}\mu_{22}(1+2\mu_{21})$	$\frac{1}{2}\mu_{22}^2$
$g_3 \otimes g_3$	μ_{21}^2	$2\mu_{21}\mu_{22}$	μ_{22}^2

The rationale used to derive the genotypic distributions in columns 2, 3 and 4 of the table was as follows. Since, by assumption, since there is no mutation from allele A_1 to A_2, each of the parents in a mating type $g_1 \otimes g_1 = A_1A_1 \otimes A_1A_1$ will produce gametes of type A_1 with probability one. Therefore, in the row of the table for this mating type, the distribution of the three genotypes g_1, g_2 and g_3 among the children is $(1,0,0)$ as indicated in the second row of the table.

For a mating of type $g_1 \otimes g_2 = A_1A_1 \otimes A_1A_2$, the parent on the right will produce gametes of types A_1 and A_2 with probabilities $\frac{1}{2}(1+\mu_{21})$ and $\frac{1}{2}\mu_{22}$, respectively. Hence, the genotypic distribution among the children of this type of mating is $\left(\frac{1}{2}(1+\mu_{21}), \frac{1}{2}\mu_{22}, 0\right)$ as indicated in the third row of the table.

With regard to a mating of type $g_1 \otimes g_3 = A_1A_1 \otimes A_2A_2$, the probability that a parent with genotype A_2A_2 produces a gamete of type A_1 is $\frac{1}{2}\mu_{21} + \frac{1}{2}\mu_{21} = \mu_{21}$. Similarly, the probability a parent of this genotype produces a gamete of type A_2 is μ_{22}. Therefore, the genotypic distribution in the fourth row of the table is $(\mu_{21}, \mu_{22}, 0)$ as indicated.

For the mating of the type $g_2 \otimes g_2$ in row 5 of the table, each parent will produce a gamete of type A_1 with probability $p = \frac{1}{2}\left(1 + \mu_{21}\right)$ and a gamete of type A_2 with probability $q = \frac{1}{2}\mu_{22}$. For this case, therefore, the genotypic distribution among the children is $\left(p^2, 2pq, q^2\right)$ as indicated in row 5 of the table. The derivations of the formulas for the genotypic distributions in rows 6 and 7 of the table may be derived by a similar rationale, but the details will be left as exercises for the reader.

The last steps in the formulation of the model consists of a set of steps in computing Monte Carlo realizations of a six-dimensional Wright-Fisher gamete sampling process with constant population size N. Suppose in generation n the state of the process is the 1×6 vector $\boldsymbol{i} = (i_1, i_2, \ldots, i_6) \in \mathfrak{S}$, and let $\boldsymbol{p}_n(\boldsymbol{i}) = \boldsymbol{i}/2N = \left(p_{n1}(i_1), p_{n2}(i_2), \ldots, p_{n6}(i_6)\right)$ denote a vector of the frequencies of the six types described in (5.5.2). Then, given $\boldsymbol{p}_n(\boldsymbol{i})$, the vector $\boldsymbol{r}_{nf}$ in (5.5.3) for females would be calculated as

$$\boldsymbol{r}_{nf} = \frac{1}{p_{n2}(i_2) + p_{n3}(i_3)}\left(p_{n2}(i_2), p_{n3}(i_3)\right). \tag{5.5.6}$$

Similarly, the vector in $\boldsymbol{r}_{nm}$ in (5.5.4) for males would be calculated as

$$\boldsymbol{r}_{nm} = \frac{1}{p_{n5}(i_5) + p_{n6}(i_6)}\left(p_{n5}(i_5), p_{n6}(i_6)\right). \tag{5.5.7}$$

After calculating the above vectors for females and males, the next step in the procedure would be that of calculating the frequencies of genotypes g_1, g_2 and g_3 as indicated in table 5.5.1. Then, given these frequencies of genotypes, the next step would be the calculation of the frequencies of the mating types in table 5.5.2. Let m_{ij} denote the frequency of matings of type $g_i \otimes g_j$.

Another component of natural selection is that of reproductive success for each of the types of matings in table 5.5.2. To measure reproductive success for a mating of type $g_i \otimes g_j$, let λ_{ij} denote the expected number of offspring produced by a mating of this type. From table 5.5.3, for matings of types $g_1 \otimes g_j$, for $j = 1, 2, 3$, it can be seen that all offspring of such matings would carry the mutant allele A_1, and, therefore, would be at risk of developing autism. Furthermore, when parents realize that one or more of their children have developed autism, they may be less likely to have more children. Based of this rationale, it would seem reasonable to assign the expectations λ_{1j}, for $j = 1, 2, 3$, smaller values than those for the other expectations λ_{ij}, for $i \neq 1$.

To simplify the notation, set up a one-to-one correspondence $(i, j) \leftrightarrow \nu$, where $\nu = 1, 2, \ldots, 6$, between mating types and a single subscript, using the ordering of the mating types in table 5.5.2. That is, $(1, 1) \rightarrow 1$,

$(1,2) \leftrightarrow 2, \ldots, (3,3) \leftrightarrow 6$. Then, let m_ν denote the frequency of mating type ν as indicated in the second column of table 5.5.2. Given this ordering of the mating types, let λ_ν, for $\nu = 1, 2, \ldots, 6$, denote the measures of reproductive success for each mating type. To take into account, the genotypic distribution for the offspring of each mating type, let $\gamma_{\nu\xi}$ denote the probability that an offspring of mating type ν is of genotype g_ξ for $\xi = 1, 2, 3$. From table 5.5.3, for example, it can be seen that $\gamma_{11} = 1$ and $\gamma_{21} = \frac{1}{2}\left(1 + \mu_{21}\right)$. Let $\boldsymbol{E} = \left(m_\nu \lambda_\nu \gamma_{\nu j}\right)$ denote a 6×3 expectation matrix, where $m_\nu \lambda_\nu \gamma_{\nu j}$ is the expected number of offspring of genotype g_j produced by matings of type ν. By way of an application of this notation, observe that the expected number of children of genotype g_1 in generation $n + 1$ produced by all the mating types in generation n is

$$\sum_{\nu=1}^{6} m_v \lambda_\nu \gamma_{\nu 1}. \tag{5.5.8}$$

The frequency of allele A_1 among the female offspring in generation $n + 1$ with symptoms of autism is proportional to the expectation

$$\eta_1 = p_f \left(\sum_{\nu=1}^{6} m_v \lambda_\nu \left(\gamma_{\nu 1} + \frac{1}{2} \gamma_{\nu 2} \right) \right) \pi_f, \tag{5.5.9}$$

and the frequency of allele A_1 among female offspring in generation $n + 1$ who do not have symptoms of autism is proportional to the expectation

$$\eta_2 = p_f \left(\sum_{\nu=1}^{6} m_v \lambda_\nu \left(\gamma_{\nu 1} + \frac{1}{2} \gamma_{\nu 2} \right) \right) \left(1 - \pi_f\right). \tag{5.5.10}$$

Similarly, it can be seen that the frequency of allele A_2 among the female offspring in generation $n + 1$ is proportional to the expectation

$$\eta_3 = p_f \left(\sum_{\nu=1}^{6} m_v \lambda_\nu \left(\frac{1}{2} \gamma_{\nu 2} + \gamma_{\nu 3} \right) \right). \tag{5.5.11}$$

Next, let the expectations η_4, η_5 and η_6 be defined for male offspring in generation $n + 1$ in ways that are completely analogous to the definitions of $\eta_\nu, \nu = 1, 2, 3$, for female offspring in generation $n + 1$, and let

$$\boldsymbol{\eta} = \left(\eta_1, \eta_2, \ldots, \eta_6\right) \tag{5.5.12}$$

denote a 1×6 vector of these expectations. Then, the frequencies of the 6 types of gametes among the offspring in generation $n + 1$ are given by the 1×6 vector

$$\boldsymbol{\zeta}_{n+1} = \frac{1}{\sum_{\nu=1}^{6} \eta_\nu} \boldsymbol{\eta}. \tag{5.5.13}$$

According to a Wright-Fisher gamete sampling process, the state

$$\boldsymbol{j} = (j_1, j_2, \ldots . j) \in \mathfrak{S} \tag{5.5.14}$$

of the process in generation $n+1$ is a realization of a random vector from a multinomial distribution with probability vector $\boldsymbol{\zeta}_{n+1}$ and index $2N$. Given the state vector $\boldsymbol{j}$, the random frequencies of the six types of gametes in generation $n+1$ is $\boldsymbol{p}_{n+1} = \boldsymbol{j}/2N$. After calculating this vector, the steps outlined above are repeated to calculate a vector $\boldsymbol{\zeta}_{n+2}$ and so the process continues for as many generations as desired. To study the evolution of autism in a population, it would be of interest to choose initial vector such that all the frequencies of gametes containing the allele A_1 are 0. For this case, the initial vector would have the form $\boldsymbol{\zeta}_0 = (0, 0, p_f, 0, 0, p_m)$, where $p_f = 100/205$ and $p_m = 105/205$.

It is interesting to note that there is a deterministic model embedded in a stochastic Wright-Fisher process. If one omits the steps in which the random vectors of frequencies, $\boldsymbol{p}_0, \boldsymbol{p}_1, \ldots$ are calculated, then one could calculate a sequence of vectors $\boldsymbol{\zeta}_0, \boldsymbol{\zeta}_1, \ldots$ in a purely deterministic way. Moreover, to assess the impact of stochasticity as formulated in a Wright-Fisher process, those two sequences could be compared in statistical summaries of a sample of Monte Carlo realizations of the Wright-Fisher process.

Among the children in any generation n, females with phenotype A_1- will present symptoms of autism with probability π_f, the penetrance of allele A_1 in females. Similarly, among males those with phenotype A_1- will present symptoms of autism with probability π_m. Among females and males, there will also be children who do not present symptoms of autism, even though some of them may be carriers of allele A_1. In any generation, the incidence of autism in children with symptoms is measured in terms of the frequencies of children with autism. Therefore, it is necessary to include procedures for calculating frequencies of children with symptoms of autism in every generation. Fortunately, some formulas presented above, may be easily modified to carry out such calculations.

For example, in any generation n, the frequency of female children who present symptoms of autism is proportional to the expectation

$$\theta_{f1} = p_f \left(\sum_{\nu=1}^{6} m_v \lambda_\nu (\gamma_{\nu 1} + \gamma_{\nu 2}) \right) \pi_f . \tag{5.5.15}$$

And the frequency of female children who do not present symptoms of autism, is proportional to the expectation

$$\theta_{f2} = p_f \left(\sum_{\nu=1}^{6} m_v \lambda_\nu (\gamma_{\nu 1} + \gamma_{\nu 2}) \right) (1 - \pi_f) + p_f \sum_{\nu=1}^{6} m_v \lambda_\nu \gamma_{\nu 3} . \tag{5.5.16}$$

Analogous expectations for male children, denoted by θ_{m1} and θ_{m2}, could be calculated by exchanging the symbols p_f and π_f with the symbols p_m and π_m in expressions (5.5.14) and (5.5.15) . Let $\boldsymbol{\theta}$ denote the vector

$$\boldsymbol{\theta} = (\theta_{f1}, \theta_{f2}, \theta_{m1}, \theta_{m2}) . \tag{5.5.17}$$

Then, in the population of children in some generation n, the frequencies of the female and male children who do and do not present symptoms of autism is given by the vector

$$\boldsymbol{\vartheta} = (\iota_1, \iota_2, \iota_3, \iota_4) = \frac{1}{\theta_{f1} + \theta_{f2} + \theta_{m1} + \theta_{m2}} \boldsymbol{\theta}. \tag{5.5.18}$$

In particular, the incidence of autism among children in generation n is

$$INC = \iota_1 + \iota_3. \tag{5.5.19}$$

At this juncture, it should be pointed out that a source of potentially high levels of stochasticity have been ignored in the above formulation. In particular, because the probability μ_{21} of mutation may be very small, the actual number of mutations observed in families may vary greatly among mating types and would certainly depend on the number of each mating type, see table 5.5.3. When these random numbers are ignored in formulation and mating types are represented only by their frequencies however, the stochastic effects in the number of mutants in the population in any generation would not be captured in summaries such as the expectations in (5.5.9) and (5.5.10) . In a subsequent chapter, stochastic models that are descendants of branching process will be formulated to model the evolution of autism in a population. In such models, the size of the population is random and may vary from generation to generation, and this randomness will be an integral part of the analysis of such processes. It should also be mentioned that the genotypic distributions in table 5.5.3 will also play a fundamental role in these stochastic models.

Mention should also be made as to the effects of stochasticity on the ratios in (5.5.13) and elsewhere. From the analysis presented in this section, it is not possible to assess the effects of randomness on these ratios. In a branching process setting however, it was shown in Mode (1985) that ratios of random functions of this form could be approximated by ratios of expectations, when population size was large.

5.6 Multitype Gamete Sampling Processes as Conditioned Branching Processes

In the multitype gamete sampling processes described so far in this chapter and chapter 4, the inclusion of selection in a formulation was accommodated in one of two ways. In section 5.2, for example, it was incorporated in the model by associating a selection probability η_ν, where $\nu = 1, 2, \ldots, k$, for each of the $k \geq 2$ types of gametes under consideration. In section 5.5 however, two facets selection were accommodated in the formulation by formally excluding females and males with autism from the mating process and associating parameters $\lambda_\nu > 0$ with each type of mating, where λ_ν was the expected number of offspring produced by couple classified as mating type ν, where $\nu = 1, 2, \ldots, 6$. As will be demonstrated in this section, when certain classes of branching processes are considered as models for the evolution of a population, selection may be accommodated in a formulation in terms of the expected number of offspring attributed to each type of individual or gamete under consideration. In this connection, many explicit results may be obtained if it is assumed that all offspring distributions are Poisson with a parameter for each type of gamete or individual under consideration. Before these explicit formulas can be derived to provide clearer insights into problem of formalizing the process of natural selection and mutation, however, it will be necessary to develop some of the properties of the Poisson distribution.

Suppose a random variable X, taking values in the set $(x \mid x = 0, 1, 2, \ldots)$, has a Poisson distribution with parameter $\lambda > 0$. Then, its probability density function is

$$f(x) = \exp(-\lambda) \frac{\lambda^x}{x!} \tag{5.6.1}$$

for $x = 0, 1, 2, \ldots$. By definition, the generating function of the random variable X is

$$\begin{aligned} G(s) = E\left[s^X\right] &= \sum_{x=0}^{\infty} f(x) s^x = \exp(-\lambda) \sum_{x=0}^{\infty} \frac{(\lambda s)^x}{x!} \\ &= \exp(-\lambda) \exp(\lambda s) = \exp(\lambda(s-1)). \end{aligned} \tag{5.6.2}$$

The generating function is also a useful tool for finding the expectation and variance of the random variable X. For by definition,

$$E[X] = \sum_{x=0}^{\infty} x f(x), \tag{5.6.3}$$

and from the observation

$$\frac{dG(s)}{ds} = \sum_{x=0}^{\infty} xf(x)\, s^{x-1}, \tag{5.6.4}$$

it can be seen that

$$\frac{dG(1)}{ds} = \sum_{x=0}^{\infty} xf(x) = E[X]. \tag{5.6.5}$$

However, $G^{'}(s) = \lambda \exp(\lambda(s-1))$ so that $E[X] = \lambda$. By a slightly more complicated argument, it can be also shown that $var[X] = \lambda$.

Generating functions are also very useful in deducing the distribution of the sum of two independent Poisson random variables. To demonstrate this statement, let the random X_1 have a Poisson distribution with parameter λ_1 and suppose the independent random variable X_2 also has a Poisson distribution with parameter λ_2. Let $f_1(x_1)$ and $f_2(x_2)$ denote the densities of these two random variables. Then, by the definition of independence, the joint density of the random variables X_1 and X_2 is

$$f(x_1, x_2) = f_1(x_1)\, f_2(x_2) \tag{5.6.6}$$

for all pairs (x_1, x_2), where $x_\nu = 0, 1, 2, \ldots$ for $\nu = 1, 2$. Now consider the random variable $Z = X_1 + X_2$. By definition, the generating function of the random variable Z is

$$\begin{aligned} G_Z(s) &= E\left[s^Z\right] = E\left[s^{X_1} s^{X_2}\right] \\ &= \sum_{x_1=0}^{\infty} \sum_{x_2=0}^{\infty} s^{x_1} s^{x_2} f_1(x_1)\, f_2(x_2) \\ &= \left(\sum_{x_1=0}^{\infty} s^{x_1} f_1(x_1)\right)\left(\sum_{x_2=0}^{\infty} s^{x_2} f_2(x_2)\right) \\ &= G_{X_1}(s)\, G_{X_2}(s). \end{aligned} \tag{5.6.7}$$

However,

$$\begin{aligned} G_{X_1}(s)\, G_{X_2}(s) &= \exp(\lambda_1(s-1)) \exp(\lambda_2(s-1)) \\ &= \exp((\lambda_1 + \lambda_2)(s-1)). \end{aligned} \tag{5.6.8}$$

Therefore, the distribution of the random variable Z is Poisson with parameter $\lambda_1 + \lambda_2$.

This result is also true for any collection of independent Poisson random variables $X_1, X_2, \ldots, X_k$ for $k \geq 2$. To state this result in a succinct symbolic form, if a random variable X has a Poisson distribution with parameter λ, then we shall write $X \sim Pois(\lambda)$. Then, a generalization of the result in (5.6.8) may be stated as follows. If $X_\nu \sim Pois(\lambda_\nu)$ for $\nu = 1, 2, \ldots, k$ and the random variables $X_1, X_2, \ldots, X_k$ are independent, then it can be shown by using generating functions that

$$Z_k = X_1 + X_2 + \cdots + X_k \sim Pois(\lambda_1 + \lambda_2 + \cdots + \lambda_k). \tag{5.6.9}$$

There is also an interesting connection between the Poisson distribution and the multinomial distribution, which is particularly useful when mutation is incorporated into a multitype gamete sampling model. Let $k \geq 2$ denote the dimension of a multinomial distribution, and let $\boldsymbol{\xi}_\nu$, for $\nu = 1, 2, \ldots, n$, be sequence of independent generalized Bernoulli indicators of dimension k. By way of an illustrative example, recall that if $k = 2$, then $\boldsymbol{\xi} = (\xi_1, \xi_2)$, where $\xi_\nu = 0$ or 1 for $\nu = 1, 2$ and $\xi_1 + \xi_2 = 1$. Let (p_1, p_2) denote a probability vector for a multinomial distribution of dimension $k = 2$. Then, by definition, the probability density function of the vector indicator $\boldsymbol{\xi}$ is

$$f_2(\xi_1, \xi_2) = p_1^{\xi_1} p_2^{\xi_2}. \tag{5.6.10}$$

Therefore, the generating function of the vector indicator $\boldsymbol{\xi}$ is

$$g(s_1, s_2) = E\left[s_1^{\xi_1} s_2^{\xi_2}\right] = p_1 s_1 + p_2 s_2. \tag{5.6.11}$$

Now consider the sum

$$\boldsymbol{Z} = \sum_{\nu=1}^{n} \boldsymbol{\xi}_v \tag{5.6.12}$$

of indicator vectors, which, as was shown in chapter 1, has a multinomial distribution with generating function $(p_1 s_1 + p_2 s_2)^n$. If $n = 0$, then the sum in (5.6.12) will be the zero vector.

A useful and interesting case arises, when it is assumed that the number n in this sum is a realization of Poisson random variable N with parameter $\lambda > 0$, and that given the realization $N = n$, the vector summands in (5.6.12) are conditionally independent. Under these assumptions, the generating function of the random vector $\boldsymbol{Z} = (Z_1, Z_2)$ is

$$\begin{aligned} G(s_1, s_2) = E\left[s_1^{Z_1} s_2^{Z_2}\right] &= \sum_{n=0}^{\infty} \exp(-\lambda) \frac{(\lambda(p_1 s_1 + p_2 s_2))^n}{n!} \\ &= \exp(-\lambda) \exp(\lambda(p_1 s_1 + p_2 s_2)) \\ &= \exp(\lambda p_1 (s_1 - 1) + \lambda p_2 (s_2 - 1)) \\ &= \exp(\lambda p_1 (s_1 - 1)) \exp(\lambda p_2 (s_2 - 1)). \end{aligned} \tag{5.6.13}$$

Therefore, $Z_1 \sim Pois\,(\lambda p_1)\,, Z_2 \sim Pois\,(\lambda p_2)$ and Z_1 and Z_2 are independent. In general, the same ideas may be used to show that if $(p_1, p_2, \ldots, p_k)$ is a vector of probabilities for a multinomial distribution of dimension $k \geq 2$, then the random variables in the vector sum $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_k)$ have the properties that $Z_\nu \sim Pois\,(\lambda p_\nu)$ for $\nu = 1, 2, \ldots, k$ and they are independent.

There is also another connection between a set of independent Poisson random variables and a multinomial distribution. For the case, $k = 2$, let $X_\nu \sim Pois\,(\lambda_\nu)$ for $\nu = 1, 2$ and suppose the random variables X_1 and X_2 are independent. Then,

$$X_1 + X_2 \sim Pois\,(\lambda_1 + \lambda_2)\,, \tag{5.6.14}$$

and the conditional density of X_1 and X_2, given that $X_1 + X_2 = N$, is, by definition,

$$\begin{aligned} f\,(x_1, x_2 \mid X_1 + X_2 = N) &= \frac{\exp\left(-\left(\lambda_1 + \lambda_2\right)\right) \frac{\lambda_1^{x_1}}{x_1!} \frac{\lambda_2^{x_2}}{x_2!}}{\exp\left(-\left(\lambda_1 + \lambda_2\right)\right) \frac{(\lambda_1 + \lambda_2)^N}{N!}} \\ &= \frac{N!}{x_1! x_2!} \frac{\lambda_1^{x_1} \lambda_2^{x_2}}{\left(\lambda_1 + \lambda_2\right)^N}. \end{aligned} \tag{5.6.15}$$

From this formula, it can be seen that the conditional distribution of the random variables X_1 and X_2, given that $X_1 + X_2 = N$, is a two dimensional multinomial distribution with index N and probability vector

$$\boldsymbol{p} = \frac{1}{\lambda_1 + \lambda_2}\,(\lambda_1, \lambda_2)\,. \tag{5.6.16}$$

In general, for the case of $k \geq 2$ independent random variables such that $X_\nu \sim Pois\,(\lambda_\nu)$ for $\nu = 1, 2, \ldots, k$, it can be shown that the conditional distribution of $X_1, X_2, \ldots, X_k$, given that $X_1 + X_2 + \cdots + X_k = N$, is a k dimension multinomial with index N and probability vector

$$\boldsymbol{p} = \frac{1}{\lambda_1 + \lambda_2 + \cdots + \lambda_k}\,(\lambda_1, \lambda_2, \ldots, \lambda_k)\,. \tag{5.6.17}$$

Having developed some properties of the Poisson distribution, we are now in a position to apply them in genetics. Suppose that in some generation n the state of the population is given by the vector $\boldsymbol{i} = (i_1, i_2)$ such the $i_1 + i_2 = 2N$, where N is the size of the population under consideration. As a first step in applying these results described above to a gamete sampling process with $k = 2$, suppose that each of the i_1 individuals in generation n of type 1 produces gametes independently according to a Poisson distribution with parameter λ_1, and, similarly, suppose each of the i_2 individuals of type 2 produces gametes independently according to a Poisson distribution with parameter λ_2. Furthermore, suppose all individuals of type 1 and 2 produce gametes independently. Then, by applying a result stated above, it follows that the total number of gametes produced by each of the two types of individuals have independent Poisson distributions with parameters $i_1\lambda_1$ and $i_2\lambda_2$, respectively.

Let

$$\mathfrak{M} = \begin{pmatrix} \mu_{11} & \mu_{12} \\ \mu_{21} & \mu_{22} \end{pmatrix} \tag{5.6.18}$$

denote the matrix of mutation probabilities, and let the random variable $Z_{\nu\nu'}$ denote the numbers of gametes of type ν' produced by the individuals of type ν. Then, the set of four independent random variables $(Z_{\nu\nu'} \mid \nu = 1, 2; \nu' = 1, 2)$ may be represented in the 2×2 matrix

$$\begin{pmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{pmatrix}. \tag{5.6.19}$$

According to the theory outlined above, the four random variables in this matrix are independent and have Poisson distributions with parameters as indicated in the matrix

$$\begin{pmatrix} i_1\lambda_1\mu_{11} & i_1\lambda_1\mu_{12} \\ i_2\lambda_2\mu_{21} & i_2\lambda_2\mu_{22} \end{pmatrix}. \tag{5.6.20}$$

With reference to the matrix (5.6.19), the total number of gametes produced by individuals of type 1 is given by the sum $Z_{11} + Z_{12}$, which has a Poisson distribution with parameter

$$i_1\lambda_1\mu_{11} + i_1\lambda_1\mu_{12} = i_1\lambda_1 (\mu_{11} + \mu_{12}) = i_1\lambda_1. \tag{5.6.21}$$

Similarly, one may conclude that the total number of gametes produced by individuals of type 2 is given by the sum $Z_{21} + Z_{22}$, which has a Poisson distribution with parameter $i_2\lambda_2$. Furthermore, the total number of gametes of type 1 in the gametic pool is given by the random variable $Z_{11} + Z_{21}$, which has a Poisson distribution with parameter $i_1\lambda_1\mu_{11}+i_2\lambda_2\mu_{21} = \gamma_1 (\boldsymbol{i})$. Similarly, the total number of gametes in the gamete pool of type 2 is given by the random variable $Z_{12} + Z_{22}$, which has a Poisson distribution with parameter $i_1\lambda_1\mu_{12} + i_2\lambda_2\mu_{22} = \gamma_2 (\boldsymbol{i})$ Finally, the total number of gametes in the gamete pool is given by the random variable

$$\sum_{\nu=1}^{2} \sum_{\nu'=1}^{2} Z_{\nu\nu'}, \tag{5.6.22}$$

which has a Poisson distribution with parameter $i_1\lambda_1 + i_2\lambda_2 = \gamma (\boldsymbol{i})$.

From the results outlined above, it may be concluded that, given the number of gametes chosen to constitute generation $n+1$ must be $2N$, it follows that the conditional distribution of the random variable $Z_{11}+Z_{21}$ and $Z_{12}+Z_{22}$, given that their sum is $2N$, is multinomial with index $2N$ and probability vector

$$\boldsymbol{p}\left(\boldsymbol{i}\right)=\frac{1}{\gamma\left(\boldsymbol{i}\right)}\left(\gamma_{1}\left(\boldsymbol{i}\right),\gamma_{2}\left(\boldsymbol{i}\right)\right). \tag{5.6.23}$$

Therefore, the number of individuals of each type in generation $n+1$ is given by the random vector $\boldsymbol{j}=\left(j_{1},j_{2}\right)$, which is a realization from the multinomial distribution described in $\left(5.6.23\right)$.

These results for the case $k=2$ may be easily extended to cases $k>2$. Let the vector $\boldsymbol{i}=\left(i_{1},i_{2},\ldots,i_{k}\right)$, where $i_{1}+i_{2}+\cdots+i_{k}=2N$, denote the state of the population in some generation n, suppose the k types of individual produce gametes independently according to Poisson distributions with parameters λ_{ν} for $\nu=1,2,\ldots,k$ and let

$$\mathfrak{M}=\left(\mu_{\nu\nu'}\right) \tag{5.6.24}$$

denote a $k\times k$ matrix of mutation probabilities. Then, let the random variable $Z_{\nu\nu'}$ denote the number of gametes produced by the i_{ν} individuals of type ν that are of type ν', where $\nu,\nu'=1,2,\ldots,k$, and let the $k\times k$ matrix

$$\left(Z_{\nu\nu'}\right) \tag{5.6.25}$$

denote the array of these random variables. According to the properties of the Poisson distribution outlined above and the assumptions of independence, all random variables in this matrix are independent with Poisson distributions with parameters as indicated in the $k\times k$ matrix

$$\left(i_{\nu}\lambda_{\nu}\mu_{\nu\nu'}\right). \tag{5.6.26}$$

From this matrix, it can be seen that for each $\nu=1,2,\ldots,k$, the distribution of the random variable

$$\sum_{\nu'=1}^{k}Z_{\nu\nu'} \tag{5.6.27}$$

is Poisson with parameter

$$i_{\nu}\lambda_{\nu}\sum_{\nu'=1}^{k}\mu_{\nu\nu'}=i_{\nu}\lambda_{\nu}. \tag{5.6.28}$$

Therefore, the distribution of the total number of gametes in the gamete pool of the population is given by the random variable

$$Z = \sum_{\nu=1}^{k} \sum_{\nu'=1}^{k} Z_{\nu\nu'}, \tag{5.6.29}$$

which has a Poisson distribution with parameter

$$\gamma\left(\boldsymbol{i}\right) = \sum_{\nu=1}^{k} i_{\nu}\lambda_{\nu}. \tag{5.6.30}$$

In a similar vein, the number of gametes of type ν' in the population is given by the random variable

$$Z_{\nu'} = \sum_{\nu=1}^{k} Z_{\nu\nu'}, \tag{5.6.31}$$

which has a Poisson distribution with parameter

$$\gamma_{\nu'}\left(\boldsymbol{i}\right) = \sum_{\nu=1}^{k} i_{\nu}\lambda_{\nu}\mu_{\nu\nu'} \tag{5.6.32}$$

for $\nu' = 1, 2, \ldots, k$. Let the random vector $\boldsymbol{j} = (j_1.j_2, \ldots, j_k)$ be the state of the population in generation $n+1$ of gamete sampling process, given that the total number of gametes selected for this generation is $2N$. Then, the vector $\boldsymbol{j}$ is a realization from a multinomial distribution with index $2N$ and probability vector

$$\boldsymbol{p}\left(\boldsymbol{i}\right) = \frac{1}{\gamma\left(\boldsymbol{i}\right)}\left(\gamma_1\left(\boldsymbol{i}\right), \gamma_2\left(\boldsymbol{i}\right), \ldots, \gamma_k\left(\boldsymbol{i}\right)\right). \tag{5.6.33}$$

It is interesting to observe that if $\lambda_1 = \lambda_2 = \cdots = \lambda_k = \lambda$ and if $\mu_{\nu\nu'} = 0$ if $\nu' \neq \nu$ so that there is no selection and mutation, then the model under consideration reduces a neutral Wright-Fisher model.

The result in (5.6.33) may be easily extended to take into account the effects of the process of immigration into a population, a process that is thought to have had significant impacts on the evolution of many populations. Let the vector,

$$\boldsymbol{W}_n = (W_{1n}, W_{2n}, \ldots, W_{kn}) \tag{5.6.34}$$

denote the random numbers of each of the k types, who immigrate into a population in generation n, and suppose the generating function of the distribution of this random vector is

$$g\left(s_1, s_2, \ldots, s_k\right) = \exp\left(\sum_{\nu=1}^{k} \alpha_k \left(s_k - 1\right)\right) \tag{5.6.35}$$

for every generation $n \geq 0$. From this formula, it can be seen that it has been assumed that all the random variables in the vector $\boldsymbol{W}_n$ are independent and have Poisson distributions with parameters $\alpha_\nu > 0$ for $\nu = 1, 2, \ldots, k$. Furthermore, suppose that for every generation n the random variables in the vector $\boldsymbol{W}_n$ are independent of the random variables in the matrix displayed in (5.6.25). Then, by using the arguments outlined above, if the state of the population in generation n is given by the random vector $\boldsymbol{i} = (i_1, i_2, \ldots, i_k)$, then the state of the population in generation $n+1$ is given by the vector $\boldsymbol{j} = (j_1, j_2, \ldots, j_k)$, which is a realization from a multinomial distribution with index $2N$ and probability vector

$$\boldsymbol{p}(\boldsymbol{i}) = \frac{1}{\gamma(\boldsymbol{i}) + \alpha} \left(\gamma_1(\boldsymbol{i}) + \alpha_1, \gamma_2(\boldsymbol{i}) + \alpha_2, \ldots, \gamma_k(\boldsymbol{i}) + \alpha_k\right), \qquad (5.6.36)$$

where $\alpha = \alpha_1 + \alpha_2 + \cdots + \alpha_k$. The paper by Karlin and McGregor (1964) may be consulted for further details on gamete sampling processes accommodating immigration and their diffusion approximations.

The word, branching, in the title of this section stems from the observation that if the condition of restricting the total number of gametes to $2N$ in each generation is omitted, then the resulting process would be a multitype branching process in discrete time with Poisson offspring distributions accommodating mutations among the types of individuals. Such processes are often referred to multitype Galton-Watson processes, and generalizations of this class of processes will be considered in subsequent chapters.

A very interesting paper on the types of model considered in this section is that of Karlin and McGregor (1965). Of particular interest in this paper is the derivation for formulas characterizing not only the eigenvalues of the Markov transitions matrices of conditioned branching process but also formulas for their left and right eigenvectors. No algorithms for computing eigenvalues and eigenvectors were included in this paper. Some of the Markov transition matrices for the Wright-Fisher process considered in this and chapter 4 are special cases of the general structures described by Karlin and McGregor. The material in these papers deserves more attention, but, because the focus of attention in this chapter will be on Monte Carlo implementation of the classes of models under consideration, no further attention will be given theories of Markov chains that are derived from conditioning branching processes.

If a reader is uncomfortable with the assumption that all offspring distributions are Poisson, then numerical versions of other offspring distributions could be used in Monte Carlo implementations of the class of process under consideration, but the details of setting up such implementations will be left to the reader as an exercises.

5.7 On the Orderly Pursuit of Randomness Underlying Monte Carlo Simulation Methods

As indicated in chapter 1 section 1.11, a basic idea underlying all Monte Carlo simulation procedures is to compute a sequence of realizations of random variables U_ν for $\nu = 1, 2, \ldots$ with values in the interval $(0, 1)$ such that are, approximately, independently and identically distributed $(i.i.d)$. The mathematics underlying computation of such numbers with properties of randomness is part of number theory and abstract algebra. If a reader is interested in these subjects, it is suggested that the classic book, Birkhoff and Mac Lane (1953), the more recent book Goldstein (1973) or the book on number theory Vinogradov (1954) be consulted. By way of a definition, the set of positive integers consists of the whole numbers $n = 1, 2, 3, \ldots$ and the set of all integers includes zero 0 and the negative and positive integers $\pm 1, \pm 2, \pm 3, \ldots$. A result known as the Euclidean division algorithm is the statement that for any two integers a and m if we divide a by m, one gets the result $a/m = q + r/m$, a quotient q and a remainder r/m. Equivalently,

$$a = mq + r, \tag{5.7.1}$$

where r is an integer such that $0 \leq r < m$. More precisely, $r = 0, 1, 2, \ldots,$ $m - 1$.

Equation (5.7.1) also gives rise to the relation of congruence between two integers a and b. These integers are said to be congruent modulo m if there is a integer r such that $0 \leq r < m$ and

$$a = q_1 m + r$$

and

$$b = q_2 m + r. \tag{5.7.2}$$

When these equations hold, we will write $a = b \bmod m$. It is of interest to note that $a = b \bmod m$ if, and only if, the difference $a - b$ is a multiple of m. It is also interesting to note that the relation of congruence is encountered in every day life. We are all aware of the sets of positive even and odd integers, $(2n \mid n = 1, 2, \ldots)$ and $(2n + 1 \mid n = 0, 1, 2, \ldots)$. From this example, it can be seen that for every even integer $a = 0 \bmod 2$ and for every odd integer $b = 1 \bmod 2$. Similarly, our basis measuring time split into 12 hour periods for morning and afternoons is equivalent to doing arithmetic modulo 12. As will be illustrated in what follows most random number generators that have been implemented in computers are based on doing sequential arithmetic mod m, where m is some large prime number.

A history of the quest for randomness may be found in the influential book Knuth (1969) and more recent editions along with descriptions of statistical tests for *i.i.d* uniform random variables on the interval $(0, 1)$. One of the most widely used class of random number generators is multiplicative, linear congruential generators of the form

$$x_i = Bx_{i-1} \bmod m, \tag{5.7.3}$$

where x_0 is an assigned initial number in the set $(1, 2, \ldots, m-1)$ and m is a large prime. A problem that arises in designing a generator of this type is the choice of the constant B such that a computed sequence $(x_i \mid i = 0, 1, 2, \ldots, n)$ will attain the maximum period, which is $m - 1$. A generator in this class that has been implemented in many computer software packages for computers with 32 bit words is

$$x_i = 16807x_{i-1} \bmod m, \tag{5.7.4}$$

where m is the prime number $2^{31} - 1$. Observe that $7^5 = 16\,807$.

A list of factorizations of integers of the form $2^x - 1$ for integers $x = 1, 2, \ldots, 64$ may be found in Knuth and from this list one can see that $2^{31} - 1 = 2,147,483,647$ is prime. Moreover, for this generator, it can be shown that the maximum period of $2^{31} - 2 = 2,147,483,646$ is attained. To obtain a sequence of numbers in the interval (0.1), the sequence

$$(u_i \mid i = 1, 2, \ldots, n) = (x_i/m \mid i = 0, 1, 2, \ldots, n) \tag{5.7.5}$$

is computed.

The random number generator in (5.7.5) will pass several statistical tests for *i.i.d.* uniform random variables on the interval $(0, 1)$. If one plots pairs of points (u_i, u_{i+1}) for $i = 1, 2, \ldots, n-1$ however, one can see there is a tendency for points to cluster on parallel straight lines. An illustration of this property for this multiplicative linear congruential generator has been provided by Deng and Lin (2000) in Figure 1 of their paper. If the sequence in (5.7.5) was indeed a realization of *i.i.d* uniform random variables on the interval $(0, 1)$, then scatter plots of points would be expected to occur uniformly in sub-regions of a two dimensional unit square. When these authors confined attention to points (u_i, u_{i+1}) in the interval $[0.70, 0.71]$, then it appeared that the points were uniformly distributed in a scatter plot. However, if attention were focused on a shorter interval $[0.700, 0.701]$, then the points tended to fall in straight lines, which is contrary to what would be expected of the sequences of points in (5.7.5) if they were indeed a realization of *i.i.d* uniform random variables on the unit interval. Similar plots may be found in more recent editions of the book by Knuth.

The random number generator in (5.7.5) is still being used in many computer software packages even though it is generally known that the generator will fail some test for randomness. The justifications for its use are that it is easy to implement with fast speeds of execution, and if one is interested in only the one-dimensional distributions of a stochastic process, its performance is quite adequate. However, if one uses this generator to evaluate multidimensional integrals numerically, the results can be misleading.

As the execution speeds of desktop computers increase, other problems connected with the period of the random number generator in (5.7.5) arise. For example, suppose an investigator wishes to do a Monte Carlo simulation experiment that involves computing a large number of realizations of *i.i.d.* Bernoulli indicators, which are defined as follows. Let p denote a probability in $(0, 1)$ and let U denote a uniform random variable on this interval. If $U \leq p$, let the Bernoulli indicator $\xi = 1$, and if $U > p$, let $\xi = 0$. Let $(U_i \mid i = 1, 2, \ldots, N)$ denote a large sample of *i.i.d* uniform random interval on the unit interval and let $(\xi_i \mid i = 1, 2, \ldots, N)$ be a sequence of *i.i.d* Bernoulli indicators with fixed parameter $p \in (0, 1)$, then, as is well known and was illustrated in chapter 1, the random variable

$$X = \sum_{i=1}^{N} \xi_i \tag{5.7.6}$$

has a binomial distribution with index or index size N and probability p. Such a random variables arise in the class of Wright-Fisher models and other gamete sampling model that have been very influential in genetics and been described in this and chapter 4.

Now suppose one wishes to carry out a Monte Carlo simulation experiment with a Wright-Fisher model such that $N = 20,000$, the number of generations is $2,000$, and the number of Monte Carlo replications of the experiment is 100. If one uses the representation in (5.7.6) , then it would be necessary to compute $100 \times 20,000 \times 2,000 = 4,000,\ 000,\ 000$ *i.i.d.* realizations of Bernoulli indicators. By comparing this number with the period of the generator in (5.7.5) , it can be seen that

$$\frac{4,000,\ 000,\ 000}{2^{31} - 2} \simeq 1.862\,64. \tag{5.7.7}$$

Thus, in such an experiment, the number of calls to a uniform random number generator would exceed its period by a factor of nearly 2. When working with the binomial distribution it is sometimes of interest to perform such experiments to provide bench marks for the performance of the Poisson and normal approximations to the binomial distribution. In view of this result it seems advisable to work with a random number generator with a much longer period as well as a capability for providing uniformly distributed points in two and higher dimensional scatter plots.

A method that has been used extensively in designing random number generators with large period is that of multiple, linear recursive, congruential generators of the form

$$x_i = (\alpha_1 x_{i-1} + \alpha_2 x_{i2} + \cdots + \alpha_k x_{i-k}) \bmod m, \qquad (5.7.8)$$

where $i \geq k \geq 1$, m is a large prime, the multiplicative constants $\alpha_1, \alpha_2, \ldots, \alpha_k$ are assigned numbers and $x_0, x_1, \ldots, x_k$ are assigned initial integers, and, moreover, many of them may be zero. When m is prime, the theory of finite fields may be applied to find $\alpha's$ such that the sequence generated by equation (5.7.8) has a period of $m^k - 1$. The $\alpha's$ have the desired property if, and only if, the polynomial

$$f(x) = x^k - \alpha_1 x^{k-1} - \cdots - a_k \qquad (5.7.9)$$

is a primitive polynomial modulo m. That is, it has a root that is a primitive element of the field with m^k elements. The technical details will be omitted here, but an interested reader may consult the book by Knuth for details and the paper by Deng and Lin for additional references.

An application of the theory just outlined, led Deng and Lin to propose the generator for computers with 32 bit words of the form

$$x_i = (39613 x_{i-2} - x_{i-1}) \bmod m \qquad (5.7.10)$$

for $i \geq 2$, where the prime $m = 2^{31} - 1$. This generator is easy to implement with the possibility of high executions speeds, and, moreover it has the maximum period $\left(2^{31} - 1\right)^2 - 1 = 4.611\,686\,014\,132\,42 \times 10^{18}$. As it turned out, this generator provided a very interesting test case for the need to apply extensive empirical statistical tests to check whether its realizations conform to the properties of *i.i.d* uniform random variables on the unit interval. When this generator was subjected to a collection of statistical tests by L'Ecuyer and Touzin (2004), it was found that it decisively failed several statistical tests. Even though the generator did not pass some tests for randomness, the paper of Deng and Lin was, nevertheless, a valuable contribution to the literature, because it demonstrated that one should not be complacent with regard to random number generators that have been implemented by software developers in many programming languages and statistical analysis software packages.

In a follow up to the paper of L'Ecuyer and Touzin, Deng (2005) provided a detailed description of the theory and numerical calculations that led to multiple recursive generators of high order. One of his suggested generators for computers with 32 bit words has been implemented and will be used throughout this book. This generator has the form

$$\begin{aligned} A &= 50283 \times x_{i-1} \bmod m \\ B &= 21034 \times x_{i-1597} \bmod m \\ x_i &= ((6379 \times A) + (4869 \times B)) \bmod m \end{aligned} \tag{5.7.11}$$

for $i > 1597$, where the prime $m = 2^{31} - 1$. As a first step for implementing this generator, one needs to compute a seed INI with 1597 elements. This generator has passed many statistical tests for randomness.

The programming language that was used to write the software applied in this book is APL 2000, and this language has the random number generator described in (5.7.4) as the default generator. Given this generator, it is easy to compute the seed INI and update it as the random number generator is called. This language also has special so called primitives for doing modular arithmetic modulo any chosen integer $m \geq 2$. Consequently, the generator in (5.7.11) may be implemented with only few lines of code. APL 2000 is an interpreted language so that if one wishes to include many loops in a program, the execution time is very slow. Fortunately, in APL there is an "each" operator, which involves operating on nested arrays, such that, when applied in Monte Carlo simulation programs, many loops can be executed in acceptable times. It is of interest to note in closing this section that the period of the generator in (5.7.11) is very large and is given by the formula

$$\left(2^{31} - 1\right)^{1597} - 1 \simeq 1.235\,933\,243\,283\,58 \times 10^{14903}. \tag{5.7.12}$$

A reference that is still useful for an account of generating realizations random variables with well known distributions from sequences of uniform *i.i.d.* random variables is that of Newman and Odell (1971).

5.8 Design of Software and Statistical Summarization Procedures

The mathematical structure of a class of Wright-Fisher gamete sampling models with mutation, selection and two or more alleles has been given in chapter 4 and the preceding sections of this chapter. In the spirit of scientific openness, the purpose of this section is to provide a brief account of the algorithms used to simulate realizations of random variables with binomial distributions. As in a previous section, let $N \geq 0$ be an integer valued index or sample size of a binomial distribution with probability $p \in [0,1] = [x \mid 0 \leq x \leq 1]$, and let X denote a random variable with this distribution with values in the set

$$\mathfrak{R}_X = [x \mid x = 0, 1, 2, \ldots, N]. \tag{5.8.1}$$

The purest approach to simulating realizations of a binomial random variable is to use its representation as an *i.i.d.* sum of N Bernoulli indicators as was indicated in chapter 1. When N is large however, the computation of a large number of Bernoulli indicators necessitate the calling large numbers uniform random numbers on the interval $(0,1)$, which is often impractical for many desktop computers. It turns out, however, that a large number of calls to a random number generator can be reduced to just a few by using known properties of binomial random variable when N is large. As was shown in chapter 1, if p is small and N is large, the distribution of the random variable X may be well approximated by a Poisson distribution with parameter $E[X] = Np$, the expectation of X. Small values of p arise when a model, dealing with the emergence of a new mutation in a population, is under consideration. It is also well known that if N is large, then for any value of $p \in (0,1)$ the random variable X is approximately normally distributed with expectation $\mu = Np$ and variance $\sigma^2 = Np(1-p)$. In symbols, $X \simeq \mathfrak{N}(\mu, \sigma^2)$.

These two well known results were used as a guide to construct an efficient algorithm for computing realizations of binomial random variables by putting branch conditions into the computer code based on values of the expectation $E[X]$, which were easy to compute repeatedly in any simulation experiment. If, for example, $0 < E[X] \leq 15$, then the Poisson distribution with parameter Np was used to compute a realization of $X \in \mathfrak{R}_X$. If $E[X] > 15$, however, a normal approximation was used. More precisely, if $Z \simeq N(0,1)$, then the random variable $Y = Np + \sqrt{Np(1-p)}Z \simeq \mathfrak{N}(Np, Np(1-p))$. In general, because the range of the random variable Y is the set real numbers $[x \mid -\infty < x < \infty]$, an adjustment was necessary to assure that all simulated values of $X \in \mathfrak{R}_X$. Let $[Y]$ denote the greatest integer in Y. Then, by letting

$$X = \min\{[Y], N\} \tag{5.8.2}$$

will assure that a realization X will never exceed N. But, a realization of Y may also be negative. By trial and error, it was observed that if $E[X]$ were chosen sufficiently large, say $E[X] > 15$, then negative values of Y could be avoided with high probability, which provided an empirical basis for the choice of 15 as a branch point to signal the use of the Poisson or normal approximations to the binomial distribution in a computer program. To eliminate the possibility of negative values of a simulated realization of X will always be in the set $\mathfrak{R}_X$, one could also put a check in a program to compute $X = \max\{0, [Y]\}$ if $[Y] < 0$.

There are other cases that may arise when simulating realizations of a binomial random variable, particularly for the case of simulating a sample from a multinomial distribution. For example, if $N > 1$ and $p = 0$, then let $X = 0$, but if $p = 1$, the let $X = N$. Finally, if $N = 0$, then let $X = 0$. All these conditions are easy to include in a computer program, but any further details will be omitted, because they would depend on the programming language used to implement these ideas. In general however, the computation of realizations of binomial random variables reduce to programs for computing realizations of either Poisson or standard normal random variables $Z \simeq \mathfrak{N}(0,1)$, which are embedded in many software packages. Consequently, there is no need here to describe procedures of transforming *i.i.d.* uniform random variables on the interval $(0,1)$ into realizations of Poisson or standard normal random variables. It is of interest to note that in the APL code used in the Monte Carlo simulations reported in this book, computing a realization of a Poisson random variable required only one call to a uniform random number generator; whereas computing a realization of a standard normal random variable required two calls to the generator. In the computer code used in the computer experiments reported in this book, the well known Box - Muller algorithm was used to transform uniform random numbers into standard normals. A formal account of the Box - Muller algorithm may be found in Newman and Odell (1971).

The sort of evolutionary questions that attempts will be made to answer in terms of a Monte Carlo simulation experiments using Wright-Fisher gamete sampling processes are usually of the following form. Given some initial conditions, where will the population be after G generations of evolution, taking into account variability or uncertainty among realizations of the process as simulated as samples from multinomial distributions. In many of the experiments reported in this chapter, the number G will be in the thousands. To take into account the variability, it is necessary to compute samples of realizations of the process for G generations. Let $\mathfrak{X} = (x_{ij})$

denote an array of $R \geq 1$ replications or realizations of a process for G generations. For example, x_{ij} denotes a realized value of the process in the *j-th* generation of the *i-th* replication, where $i = 1, 2, \ldots, R$ and $j = 1, 2, \ldots, G$. A very useful way of answering the question where will a population be after some number of generations, given some initial conditions, is to summarize of sample of Monte Carlo realizations of a process in terms of estimated quantile trajectories.

To estimate the quantile trajectories from a Monte Carlo sample of realizations, it is necessary to order the columns of the $R \times G$ array $\mathfrak{X}$ from the largest to the smallest in each column. Let

$$\mathfrak{X}_o = \left(x_{ij}^{(o)} \right) \tag{5.8.3}$$

denote the array of ordered columns. For example, if $i = 1$, then $x_{1j}^{(0)}$ would be the largest realized value of the process in generation j in a sample of R realizations of G generations, and, similarly, if $i = R$, then $x_{Rj}^{(o)}$ would be the smallest value of the process in generation j. An estimate of the 50 percent quantile trajectory of the process would be an array of G numbers

$$Q50 = (q(50)_j \mid j = 1, 2, \ldots, G) \tag{5.8.4}$$

such at in every generation 1/2 the realizations of the process in the ordered array $\mathfrak{X}_o$ are $\leq q(50)_j$ for every $j = 1, 2, \ldots, G$. Given the ordered array $\mathfrak{X}_o$, it is easy to program a computer to select the appropriate row of $\mathfrak{X}_o$ to find a quantile trajectory. For example, if R is even, then the row with index $R/2$ would correspond to the $Q50$ trajectory. The quantile trajectories $Q25$ and $Q75$ are defined and selected similarly. Another useful trajectory for statistically summarizing a sample of Monte Carlo realizations of a process is the MAX array defined by

$$MAX = \left(x_{1j}^{(0)} \mid j = 1, 2, \ldots, G \right). \tag{5.8.5}$$

The MIN array of a sample $\mathfrak{X}$ of realizations is defined similarly. The mean ,$MEAN$, and standard deviation SD trajectories could also be computed from the simulated data.

In many cases of interest, a useful signature for the evolution of a process over a large number of generations, given some initial conditions, would be the plot of the trajectories

$$\begin{bmatrix} MAX \\ Q25 \\ Q50 \\ Q75 \\ MIN \end{bmatrix} \tag{5.8.6}$$

as a function of the number of generations $j = 1, 2, \ldots, G$. It may also be of interest to include the trajectories, $MEAN$ and SD, in graphs visually summarizing the simulated data. In some selected examples in the sections that follow, illustrative examples of such plots will be presented and discussed from the perspective of signatures for evolutionary processes. In the spirit of scientific openness in reporting the results of Monte Carlo simulation experiments, the foregoing description seems adequate.

5.9 Experiments in the Quantification of Ideas for the Evolution of Inherited Autism in Populations

The purpose of this section is to report the results some computer experiments with the evolutionary model of autism described in section 5.5 in which attempts were made to quantify some ideas for the evolution of autism in populations. Software was written to implement both the deterministic and stochastic versions of this model. As has been widely reported in scientific sources and in news items, the incidence of autism is about 1 in 150 births in the United States. To accommodate such widely used measures of incidence, in the experiments reported in this section incidence was computed as $1/INC$, where INC was defined in equation (5.5.19). Throughout all experiments reported in this section the initial frequencies in the initial 1×6 frequency vector were $\boldsymbol{p}_0 = (0, 0, p_f, 0, 0, p_m)$, indicating that in the initial population there were no individuals carrying the mutant allele A_1. The values of p_f and p_m were chosen as $p_f = 100/205$ and $p_m = 105/205$, which are used widely in the literature on human biology.

In preliminary experiments with the deterministic model, in which all steps requiring the computation of realizations from a multinomial distribution were omitted, it was found in several experiments that when the initial vector $\boldsymbol{p}_0$ was used, convergence to a constant vector occurred within $G = 2,000$ generations. The time taken to execute each of these experiments was just a few seconds. As an aid in searching the parameter space of the model with a goal of finding plausible values of parameters that would lead to an incidence of 1 in 150 births, the parameter vector $\boldsymbol{\lambda} = (1, 1, 1, 2, 2, 2)$ was chosen for all experiments reported in this section. Recall that this vector contains the expected number of offspring produced by the six mating types displayed in table 5.5.2 and provides quantification of an idea as to how parents may not want more children after that have had one autistic child. Given these assignments of parameters, there remained only three parameters, the mutation probability μ and the penetrance probabilities π_f and π_m for females and males, to consider in searches for plausible combinations of numerical values that would lead to an incidence rate of 1 in 150 births after 2,000 generations of evolution. From now on, the phrase, 1 in some number, will be referred to as the incidence rate.

Among many experiments with the deterministic model, it was shown that if $\mu = 10^{-6}, \pi_f = 0.5002$ and $\pi_m = 0.9$, then after 2,000 generations of evolution the incidence rate was about 150. Then, keeping at $\pi_m = 0.9$ and letting $\mu = 10^{-7}$, it was found that $\pi_j = 0.496$ yielded an incidence rate of about 152 after 2,000 generations of evolution. In a similar experiment, it was found that the combination of parameter values $\mu = 10^{-5}, \pi_f = 0.5444$ and $\pi_m = 0.9$ also yielded an incidence rate of about 151. These examples show that various combinations of the parameters in the triple (μ, π_f, π_m) would yield incidence rates near 150, and that if π_m was fixed, then it was possible to adjust values of π_f such that, for a given mutation probability μ, one could get incidence rates in a neighborhood of 150. By choosing values of μ for some fixed pairs of values (π_f, π_m) is was also demonstrated that incidence rates could vary greatly as a function of μ, indicating that the response of the deterministic model was also very sensitive to changes in probabilities of mutation. Evidently, incidence rates, as a function of the three parameters (μ, π_f, π_f), form some kind of complex irregular surface in four-dimensional space after 2,000 generations of evolution according to the deterministic model.

In the deterministic model, the number N of individuals in a population is not taken into account, but in the stochastic model the size N of the population is a fundamental parameter of the model. As was discussed in section 5.5, in the stochastic version of the evolutionary model of autism, in each generation $n+1$ after the frequency vector $\boldsymbol{\zeta}_{n+1}$ is calculated according to the steps outlined in section 5.5, see equation 5.5.13, then the state $\boldsymbol{j} = (j_1, j_2, \ldots, j_6)$ of the population in generation $n+1$ is a realization from a multinomial distribution with index $2N$ and probability vector $\boldsymbol{\zeta}_{n+1}$ so that the stochastic frequency vector in this generation is $\boldsymbol{p}_{n+1} = \boldsymbol{j}/2N$. As an aid to interpreting the experimental results that will be discussed subsequently, all statistical analyses will be based on $R \times G$ arrays of the form

$$\mathfrak{A} = (\boldsymbol{p}_{ij}) \tag{5.9.1}$$

where $\boldsymbol{p}_{ij}$ denote a simulated stochastic frequency vector for replication $i = 1, 2, \ldots, R$ in generation $j = 1, 2, \ldots, G$. Although several computer experiments with the stochastic model were conducted, only two, which demonstrated the importance of the demographic variable N, population size, in the evolution of incidence of autism in a population, will be reported.

In experiment 1, the number of generations of evolution was chosen as $G = 6,000$ in an attempt to ensure that the simulated data would converge in distribution to some statistical equilibrium. By way of interpreting this number, suppose that the average age of females at the birth of the first child is 15 years. Then, on average, 6,000 generations would cover $6,000 \times 15 = 90,\ 000$ years. Population size was chosen as $N = 10,000$ and the number of Monte Carlo replications of the experiment was chosen as $M = 100$. The probability of mutation was chosen as $\mu = 10^{-6}$ per generation and the penetrance of the mutant allele A_1 was chosen as $\pi_f = 0.5002$ and $\pi_m = 0.9$ for females and males, respectively. The simulated data were statistically summarized according to the procedures outlined in section 5.8 and displayed in a $6,000 \times 9$ array. The first column in this array, was the generations $1, 2, \ldots, 6,000$, and the second through ninth columns were the $Min, Q25, Q50, Q75, Max, Mean, SD$ and Det, giving the values of these statistics in each generation. The symbol, Det stands for the values produced by the deterministic model for each generation.

A visual examination of this $6,000 \times 9$ indicated that convergence to a statistical equilibrium was not occurring in the Monte Carlo data, but, as expected, the trajectory of the deterministic model did converge to a constant incidence of about 150. In fact, the level of stochasticity in the simulated Monte Carlo data was so high that plots of the summary statistics as functions of generations were not informative. A decision was made, therefore, to display a summary of the statistical data at selected generations.

Presented in table 5.9.1 are the summary statistics for $Q25, Q50, Q75$ and Det at selected generations.

Table 5.9.1 Quantile and Deterministic Trajectories of Incidence - Pop Size 10,000

G	$Q25$	$Q50$	$Q75$	Det
10	1263.070	2845.691	4885.338	1619.367
1500	244.0309	556.778	936.895	150.444
3000	217.701	340.294	819.472	150.444
4500	272.584	554.775	1560.669	150.444
6000	227.566	397.236	929.794	150.444

Table 5.9.2 contains the $Min, Max, Mean$ and SD statistics at the selected generations.

Table 5.9.2 Extreme Values, Mean and SD Trajectories of Incidence - Pop Size 10,000

G	Min	Max	$Mean$	SD
10	292.761	$16,119.892$	5714.944	6056.760
1500	67.050	$16,119.892$	1925.099	4047.626
3000	41.930	$16,119.892$	1441.870	3224.439
4500	88.237	$16,119.892$	2328.278	4502.753
6000	59.461	$16,119.892$	1514.850	3511.103

From an inspection of the fifth column in table 5.9.1, it can be seen that the trajectory of the deterministic model converged to an incidence of about 1 in a 150 births within 1,500 generations after displaying an incidence of about 1 in 1,619 births in generation 10. Interestingly, the values of incidences for the $Q50$ and $Q75$ in generation 10 are 1 in 2845 and 1 in 4885 births, respectively, indicating that the stochastic yielded lower rates of incidence than the deterministic model for this generation. There is some suggestion from the numbers for generation 6000, that the Monte Carlo data may be displaying signs of convergence in distribution, but it is of interest to note that the incidence of about 397 of the $Q50$, the median quantile, is greater than that of about 150 for the deterministic model.

From the data for the $Mean$ and SD statistics in table 5.9.2, it can be seen that the level of stochasticity remained very high throughout the table. Observe that the SD values are greater than those for the $Mean$ in every generation displayed in the table, which indicated a highly level of stochasticity in the data. The reason for this high level of stochasticity becomes apparent when the column for the Max trajectory is inspected. For from this column, it can be seen that the Max trajectory is about 16,119 for every generation displayed in the table. An inspection of the $6,000 \times 9$ array of summary statistics indicated that most of the Max trajectories over generations stood at this value but a few were less than this value.

The appearance of these values in a projection is due, in part, to the design of the model under consideration. This is the module of model taking into account the mating of females and males and the production of their offspring is purely deterministic and it has the property that if the frequency of the mutant allele A_1 is zero in some generation, then incidence in the next generation will be a constant, which is a function of the mutation parameter μ, see table 5.5.3.

For the parameter value $\mu = 10^{-6}$, this constant is about 16,119. In any generation, the values of the Max statistic is the maximum of the simulated incidences over 100 replications. Therefore, if the frequency of the mutant allele A_1 is zero in some generation in at least one of the 100 replications, then for this generation the Max statistic is about 16,119. Consequently, a value of Max at about 16,119 is an indicator of the event that the frequency of the mutant allele A_1 was zero in the gametes of the parents contributing offspring to this generation in some replication of the experiment.

As will be shown in the simulation experiment that will be described next, the reason for this propensity of Max values of about 16,119 was that a population size of only 10,000 was not sufficiently large to guarantee that with high probability the frequency of the mutant allele A_1 was positive in all replications considered in the experiment for generations $G \geq 2$.

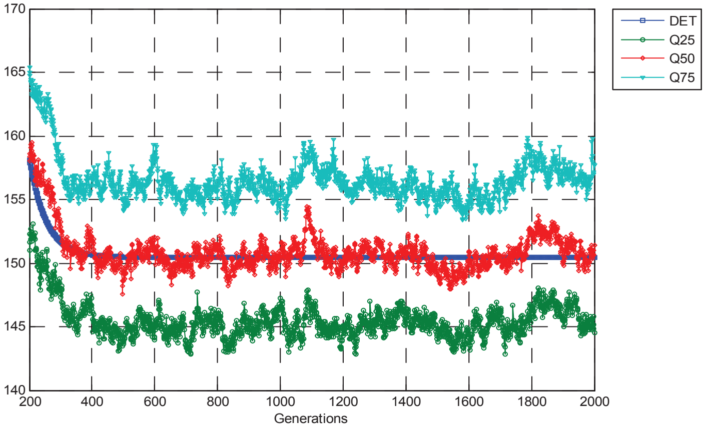

Figure 5.9.1 Graphs of the $Q25$, $Q50$, $Q75$ and Det Trajectories of the Incidence of Autism - Population Size 2,000,000.

Experiment 2 involved only two changes in the parameter values used in experiment 1. For this experiment, the number of generations of evolution considered was $G = 2,000$ and population size was chosen as $N = 2,000,000$. The reason for choosing $G = 2,000$ was that in a preliminary experiment with this population size evidence was found supporting the idea that the stochastic process was converging in distribution within this number of generations. Presented in figure 5.9.1, are the trajectories of the quantiles $Q25, Q50, Q75$ as well as the deterministic trajectory, which is labeled Det. These trajectories were plotted only for generations $n \geq 200$, because during the first 200 generations of the projections a high level of stochasticity was observed, which distorted the evidence that the process was tending to a stationary distribution in generations beyond 200.

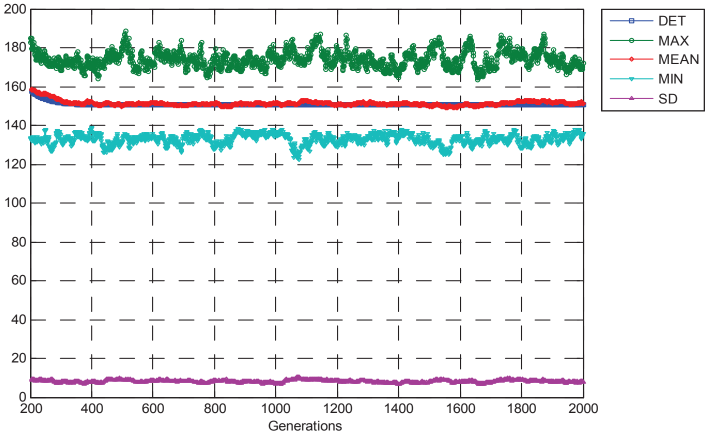

Figure 5.9.2 Graphs of the *Min*, *Max*, *Mean*, *SD* and *Det* Trajectories of the Incidence of Autism - Population Size 2,000,000.

From an inspection of the trajectories in figure 5.9.1, it appears that convergence to a stationary distribution did not occur until after the process had evolved for about 400 generations. It is also of interest to note that when the process reached an apparent statistical equilibrium, the median trajectory, $Q50$, and the deterministic trajectory were quite close for generations $n \geq 400$. In terms of calender time, if 15 is the average age for females to start childbearing, then 200 generations corresponds to $200 \times 15 = 3,000$ years and 400 generations represents $6,000$ years. Presented in figure 5.9.2 are the graphs of the $Min, Max, Mean, SD$ and Det trajectories for generations $n \geq 200$.

Upon inspection of the graphs of the $Mean$ and deterministic, Det, trajectories in this figure, it can be seen that for generations $n \geq 400$ these graphs are closer than they were for the $Q50$ trajectory in figure 5.9.1. The most striking feature of the graphs in figure 5.9.2 is that SD trajectory lies considerably below those of the other trajectories, indicating that when the process is a statistical equilibrium, the levels of variation or stochasticity among the realizations of the process is low. In computer experiments not reported here with a mutation probability of $\mu = 10^{-7}$ and the other values of the parameters set at numbers such that at equilibrium of the deterministic model the incidence rate was about 150, there was no evidence for convergence to a statistical equilibrium after 2,000 generations of evolution with a populations size of 4,000,000. When, however, the number of generations was increased to 6,000, there was evidence that the process was converging to a statistical equilibrium. Evidently, when a mutation probability is as low as 10^{-7}, then the waiting time for the process to converge to a statistical equilibrium can be in the thousand of years, which may, in some cases be longer than the period of recorded history.

5.10 Comparative Experiments in the Quantification of Two Formulations of Gamete Sampling Models

In section 5.2 a multitype gamete sampling model with mutation and selection was formulated as an extension of the two-allele Wright-Fisher processes introduced in chapter 4, and in section 5.6 multiple type gamete sampling models were introduced as conditioned branching processes, which also accommodated mutation and selection. In the model of section 5.2, selection was quantified in terms of "probabilities" that gametes were "selected" in some generation $n \geq 1$ for the gene pool of generation $n+1$. In the model introduced in section 5.6 however, selection was characterized in terms of the expected number of gametes an individual of a given type contributed to the next generation. In both models, however, the formulation of mutations among the types of gametes under consideration had the same structure. The purpose of this section is to present the results of two computer experiments in which the two ways of quantifying selection were compared. In both experiments, three types of gametes, A_1, A_2 and A_3, were considered. In terms of Mendelian genetics, these types of gametes may be interpreted as multiple alleles at some autosomal locus.

In experiment 1 the formulation of section 5.2 was considered, and the number of generations was chosen as $G = 6,000$, the number of Monte Carlo replications of the experiment was $R = 200$ and population size was $N = 2,000,000$. The selection probabilities of the three alleles were chosen as $\upsilon_1 = 0.951, \upsilon_2 = 0.952$ and $\upsilon_3 = 0.953$. Thus, in terms of selective advantage, the ranking of the three alleles was $A_3 \succ A_2 \succ A_1$ but observe the three probabilities differ in only the third place so that one would expect selection to be slow. The matrix of mutation probabilities had the form

$$\mathfrak{M} = \begin{pmatrix} \mu_{11} & 10^{-6} & 10^{-6} \\ 0 & \mu_{22} & 10^{-6} \\ 0 & 0 & 1 \end{pmatrix}, \tag{5.10.1}$$

where $\mu_{11} = 1 - 10^{-6} - 10^{-6}$ and $\mu_{22} = 1 - 10^{-6}$. Given this selection of mutation probabilities, it follows that the state $(0, 0, 2N)$ for $N = 2,000,000$ is the only absorbing state of the process and indicates that a population is homozygous for allele A_3. Finally, the initial state of the process was chosen as $(2N, 000, 0, 0)$, indicating that the population was homozygous for allele A_1. This initial state was chosen, because one of the aims of the experiments was to get some idea of the length of the waiting time to absorption in state $(0, 0, 2N)$, which according to theory would occur eventually with probability one.

In experiment 2, in which the formulation in section 5.6 was used, all parameters had the same values as in experiment 1, but the parameters quantifying the idea of selection were chosen following different fundamental principles. More precisely, λ_1, the expected number of gametes contributed by each individual of type A_1, was chosen $\lambda_1 = 2.001$. Similarly, the parameters λ_2 and λ_3 for gamete types A_2 and A_3 were chosen as $\lambda_2 = 2.002$ and $\lambda_3 = 2.003$. Just as in experiment 1, the allele A_3 has a selective advantage over the other two and the three selection parameters differ only in the third decimal place. Clearly in this experiment however, selection was characterized using different concepts than in experiment 1. Nevertheless, it is of interest to compare the results of the two experiments.

To shorten the presentation of the results of these two experiments, a decision was made to omit the graphs for the evolution of the frequencies of alleles A_1 and A_2 in the population and to confine attention only to the evolution of gene frequencies of allele A_3 with the greatest selective advantage. Briefly, in results not shown here, in both experiments the frequencies of allele A_1 decreased from an initial high of one to a very low frequency at 6,000 generations in both experiments. On the other hand, given the initial state under consideration, the frequency of allele A_2 remained low throughout the projections in both experiments. Presented in figure 5.10.1 are graphs of the $Max, Min, Q25, Q50$ and $Q75$ for the model used in experiment 1, and figure 5.10.2 contains graphs for the same trajectories for the model that was used in experiment 2.

From an inspection of the graphs in these two figures, it can be seen that the trajectory profiles of the two experiments were very similar, but it appears that in experiment 2, in which the model was based on a conditioned

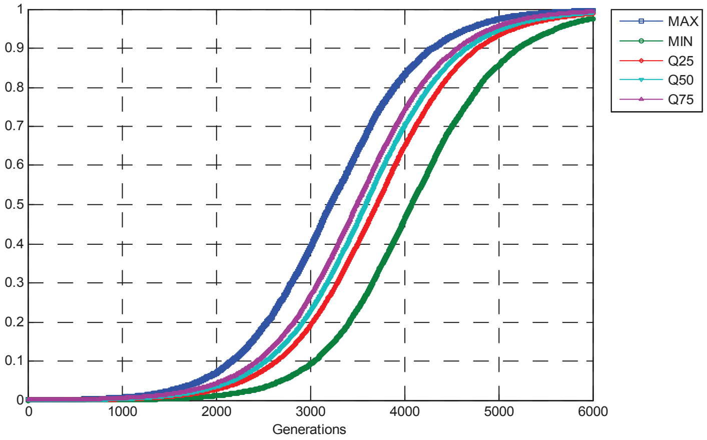

Figure 5.10.1 Graphs of Quantile and Extreme Value Trajectories of Frequencies of Allele A_3 for Experiment 1.

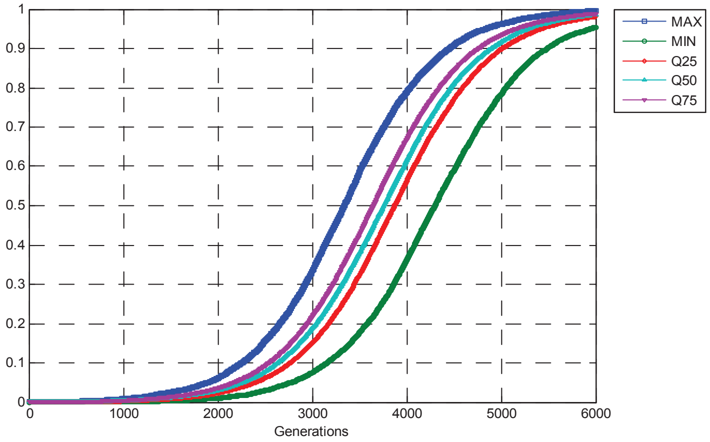

Figure 5.10.2 Graphs of Quantile and Extreme Value Trajectories of Allele A_3 Frequencies for Experiment 2.

branching process was used, the increase in the Min and Max trajectories appeared to be slower than in experiment 1. Indeed, if it were possible for a reader to superimpose the graphs and hold them up to a light, these differences could be discerned. From a substantive point of view however, these differences were judged to be insignificant. However, because selection is characterized in a more concrete way in the conditioned branching process in terms of the expected number of gametes contributed to the next generation by each type of individual, in the remaining sections of this chapter the conditioned branching process will be used as the preferred model.

5.11 An Experiment with a Three Allele Neutral Model

As was indicated in a previous section, when it is assumed that there is no selection or mutation, then the classical Wright-Fisher model with multiple alleles described in section 5.2 reduces to what is called the neutral model of evolution. In section 5.6, where a gamete sampling model was derived by conditioning a branching process on total sample size, it was shown that this model also reduces to a Wright-Fisher neutral model when there is no mutation or selection. In this section, the results of an experiment with the neutral model, using the software which implemented the model in section 5.6, will be reported. For this experiment, the number of generations considered was chosen as $G = 6,000$, the number of Monte Carlo realizations of the experiment was $R = 200$ and population size was chosen as $N = 2,000,000$. To include the condition of no selection, the lambda parameters were assigned equal values, $\lambda_1 = \lambda_2 = \lambda_3 = 2$. Actually, a particular value chosen for these parameters is of no consequence. The condition of no mutation was taken into account by setting all mutation probabilities equal to zero, *i.e.*, $\mu_{ij} = 0$ if $i \neq j$. Finally, the initial state of the population was chosen as

$$\boldsymbol{X}_0 = ((1,333,333),(1,333,333)\,,(1,333,334))\,. \tag{5.11.1}$$

Given this initial state, the probability of eventual absorption in each of the absorbing states, $(2N,0,0)$, $(0,2N,0)\,,(0,0,2N)\,,$ is $1/3$ for $N = 2,000,000$.

As the evolutionary quantile trajectory profiles over 6,000 generations of evolution were virtually the same for the three alleles, only the graphs of those for allele A_3 are presented in figure 5.11.1. From this figure it can be seen that the median trajectory, $Q50$, is not far from the initial frequency of about $1/3$ for allele A_3 throughout the projection. Moreover, at 6,000 generations, the quantiles $Q75$ and $Q25$ differ by only about 0.02 so that about 50 percent of the realizations of the process are close to the median.

Furthermore, at 6,000 generations, the difference $Max - Min$ is approximately $0.38 - 0.10 = 0.28$, indicating, as expected by the underlying theory, that there is a slow drift to fixation in one of the absorbing states.

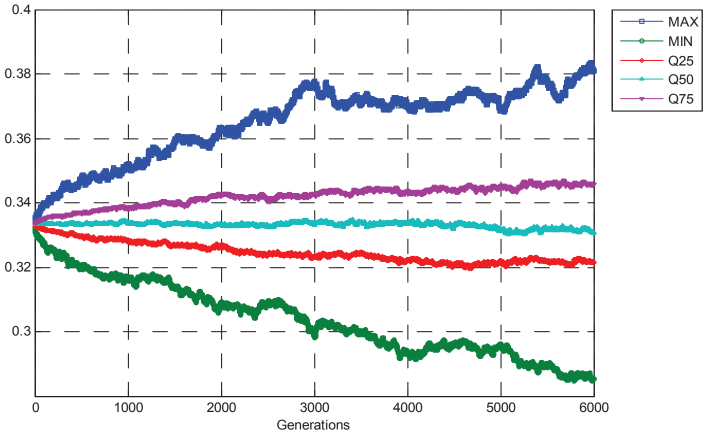

Figure 5.11.1 Quantile and Extreme Value Trajectories for Allele A_3 for a Wight-Fisher Neutral Model of Evolution with Population Size 2,000,000.

However, the values of the trajectories at 6,000 generations suggests that the quasi-stationary distribution of the process is governing the evolution of the population prior to fixation in some absorbing state at some distant future time. In this connection, it is of interest to compare the graph of the quasi-stationary distribution in figure 4.9.2 for a Wright-Fisher neutral model with population size $N = 500$ with the trajectories of gene frequencies in figure 5.11.1 at 6,000 generations.

5.12 Rapid Selection and Convergence to a Stationary Distribution

In the experiment reported in this section, the gamete sampling model described in section 5.6 was implemented for the case of three autosomal alleles with the following parameters values. The parameters G, R and N were chosen as $G = 6,000, R = 200$ and $N = 2,000,000$. The three selection parameters were chosen as $\lambda_1 = 2.1, \lambda_2 = 2.2$ and $\lambda_3 = 2.3$ so as to study a case in which these parameters differed by greater amounts than those in section 5.10, where a process of slow selection was studied. To take into account mutation, all mutation probabilities were assigned the values $\mu_{ij} = 10^{-6}$, when $i \neq j$. For these chosen values of the parameters, all states in the state space $\mathfrak{S}$ communicate and eventually the process will converge to a stationary distribution. As a test case for a rate of convergence to a stationary distribution, the initial state of the process was chosen as

$$\boldsymbol{X}_0 = (2N, 0, 0)\,, \tag{5.12.1}$$

indicating that in generation $n = 0$ the population was homozygous for allele A_1.

As expected, the frequency of allele A_1 declined rather rapidly from an initial value of 1 in generation $n = 0$ to the vary small frequency at 6,000 generations, but the frequency of allele A_2 remained low throughout the projection. On the other hand, because of its selective advantage, the frequency of allele A_3 rose quickly from a frequency of 0 in the initial generation to a very high frequency at 6,000 generations. Displayed in figure 5.12.1 are the graphs of the quantile and extreme value trajectories for the evolution of the frequency of allele A_3 for the first 100 generations of the 6,000 generations considered in the projection.

The reason for confining attention to the first 100 generations in the projection was that when the frequency of allele A_3 was graphed for 6,000 generations, the first 100 generations appeared as a mere blip on the left portions of the graphs and all trajectories rose very quickly to one, indicating that the frequencies of alleles A_1 and A_2 in the population were very small. From the graphs in figure 5.12.1 it can be seen that the frequency of allele A_3 rose to noticeable frequencies within 40 generations or within $40 \times 15 = 600$ years, if it is assumed that the average age of child bearing for females was 15 years. After 40 generations, all trajectories describing fluctuations in the frequencies of allele A_3 rose rapidly and reached frequencies very close to one at 100 generations into the projections or $100 \times 15 = 1,500$ years.

In terms of evolutionary time, 1,500 years is a very short time span and even if such rapid evolution had occurred in a population of human beings, at generation 100 it seems unlikely that most of the population would not be aware that such changes in gene frequencies had indeed occurred within the last 1,500 years, because historical data to record such an evolutionary transition has not been collected and preserved over the generations by any civilization studied in the archeological records. According to the values of the mutation parameter chosen for the experiment under consideration, it follows that the evolutionary process would converge to a stationary distribution eventually. For the case under consideration, since evolution progressed at a rapid pace, it seems plausible that the simulated statistics at 6,000 generations into the projections are a sample from the stationary

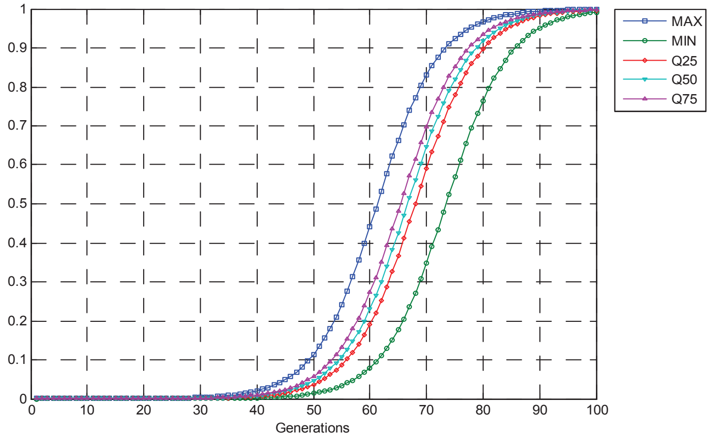

Figure 5.12.1 Quantile and Extreme Value Trajectories for A_3 for a Rapid Selection Model.

distribution of the process. Presented in table 5.12.1 are the summary statistics for each of the three alleles at 6,000 generation of the projections.

Table 5.12.1 Quantiles and Extreme Value Statistics for the Marginals of the Stationary Distribution for Each of the Three Alleles

·	A_1	A_2	A_3
Min	0.00000125	0.00000225	0.9999705
$Q25$	0.00000375	0.000007	0.999983
$Q50$	0.0000045	0.00000975	0.99998525
$Q75$	0.00000575	0.0000125	0.99998825
Max	0.0000105	0.00002175	0.999994

From the simulated data presented in this table, it can be seen that alleles A_1 and A_2 are still present in the population but at low frequencies when the population is in statistical equilibrium. Note that the $Q50$ statistic for allele A_2 is higher than that for allele A_1, which reflects the assumption that allele A_2 had a selective advantage over A_1 in the projections. From the last column of the table is can be seen that at statistical equilibrium, the distribution of the frequency of allele A_3 is tightly distributed around the median frequency $Q50 = 0.99998525$. As the numbers in the table are summary statistics based on a sample size of 200, the numbers in the rows of the table will not, in general, sum to one.

Bibliography

[1] Birkhoff, G. and MacLane, S. (1953) **A Survey of Modern Algebra**. MacMillian Company, New York.

[2] Deng, L. and Lin, K. (2000) Random Number Generation of a New Century. The American Statistician **54**:145–150.

[3] Deng, L. (2005) Efficient and portable multiple recursive generators of large order. ACM Transactions on Modeling and Computer Simulation **15**:1–13.

[4] L'Ecuyer, P. and Touzin, R. (2004) On the Deng-Lin Random Number Generator and Related Methods. Statistical Computing **14**:1–10.

[5] Goldstein, L. J. (1973) **Abstract Algebra: A First Course**. Prentice-Hall, Englewood Cliffs, New Jersey, New York.

[6] Griffiths, R. and Tarvaré, S. (1994) Simulating Probability Distributions in the Coalescence. Theoretical Population Biology. **46**:131–159.

[7] Karlin, S. and McGregor, J. (1964) Direct Product Branching Processes and Related Markov Chains. PNAS **51**:598-602.

[8] Karlin, S. and McGregor, J. (1965) Direct Product Branching Processes and Related Induced Markoff Chains. I. Calculations of Rates of Approach to Homozygosity. In Bernoulli 1713, Bayes 1765 and LaPlace 1813. Springer, New York Inc. Edited by J. Neyman and L. Le Cam. 111–145.

[9] Kennedy, W. J. and Gentle, J. E. (1980) **Statistical Computing**. Marcel Dekker, Inc. New York.

[10] Knuth, D. E. (1969) **Seminumerical Algorithms-The Art of Computer Programming**. Addison-Wesley, Reading, Mass, London, Amsterdam, Sydney.

[11] Mode, C. J. and Gallop, R. J. (2008) A Review on Monte Carlo Simulation Methods as They Apply to Mutation and Selection as Formulated in Wright-Fisher Models of Evolutionary Genetics. Mathematical Biosciences **211**:205–225.

[12] Mode, C. J. (1985) **Stochastic Processes in Demography and Their Computer Implementation**. Springer-Verlag, Berlin, Heidelberg, New York and Tokyo.

[13] Newman, T. G. and Odell, P. L. (1971) **The Generation of Random Variates**. Hafner Press, New York.

[14] Snustad, D. P. and Simmons, M. J. (2006) **Principles of Genetics**. John Wiley & Sons, Inc., New York.

[15] Strachan, T. and Read, A. (2004) **Human Molecular Genetics, Third Edition**. Garland Science, London and New York.

[16] Vinogradov, I. M. (1954) **Elements of Number Theory**. Dover Publications, Inc.

[17] Zhao, X., Leotta, A. et al. (2007) A Unified Genetic Theory for Sporadic and Inherited Autism. PNAS **104**:12831–12836.

Chapter 6

Nucleotide Substitution Models Formulated as Markov Processes in Continuous Time

6.1 Introduction

Recent advances in technology for sequencing DNA have provided insights into the mutational processes that appear to be operational at the molecular level. As was discussed in a previous chapter, among the mutational processes that the sequencing of the human genome revealed was the high prevalence of the phenomenon of nucleotide substitution. That is, at a set of sites of a DNA molecule that was under observation, a nucleotide that is present in a vast majority of individuals at a particular site has been replaced by another nucleotide at this site in some individuals. Such mutations are referred to single nucleotide polymorphisms, which are denoted by the acronym $SNP's$. This type of mutation may have a profound impact on the individuals in which they are present. For if a nucleotide substitution occurs within some codon consisting of three letters, the mutated codon may code for an amino acid that is not the same as that for the codon that is more prevalent in a population, resulting in, among other things, a possible change in the physiology of the individuals that carry the mutation. In this chapter, an overview of models that have been proposed in the literature during past forty or so years will be provided. These proposed models seem to be oversimplified versions of much more complicated and inscrutable stochastic processes that seem to be operational in nature. Virtually all the proposed models that will be discussed in this chapter are examples of Markov processes in continuous time with finite state spaces. The first item on the agenda will, therefore, will be a presentation of an overview of this class of stochastic processes, and as the chapter proceeds some discussion of the evolutionary implications of these models will also be included.

6.2 Overview of Markov Jump Processes in Continuous Time with Finite State Spaces and Stationary Laws of Evolution

Let

$$\mathfrak{S} = (i_k \mid k = 1, 2, \ldots, r) \tag{6.2.1}$$

denote the finite state space of a Markov process in continuous time with $r \geq 2$ elements, and let $X(t)$ denote a random function giving the state of the process at time $t \in [0, \infty) = T^+$. This process has the Markov property if for every $n \geq 1$, time points $t_0 < t_1 < \cdots < t_n$ in T^+ and states $(i_k \mid k = 0, 1, 2, \ldots, n)$ in $\mathfrak{S}$, the equation

$$\begin{aligned} &P\left[X(t_n) = i_n \mid X(t_k) = i_k, k = 0, 1, 2, \ldots, n-1\right] \\ &= P\left[X(t_n) = i_n \mid X(t_{n-1}) = i_{n-1}\right] \end{aligned} \tag{6.2.2}$$

holds whenever the conditional probabilities are defined. This process will be said to have stationary laws of evolution if, for all points s and t in R^+ such that for $s < t$ and states i and j in $\mathfrak{S}$, the conditional probability

$$P\left[X(t) = j \mid X(s) = i\right] \tag{6.2.3}$$

depends only on the difference $t - s$ and the states i and j For every pair of states i and j, let $P_{ij}(t)$ denote a function defined for $t \in T^+$ such that $0 \leq P_{ij}(t) \leq 1$ for all $t \in T^+$ and states i and j in $\mathfrak{S}$. Furthermore, it will be required that these functions have the property

$$\sum_{j \in \mathfrak{S}} P_{ij}(t) = 1 \tag{6.2.4}$$

for all $t \in T^+$ and states $i \in \mathfrak{S}$. Given this collection of functions, the equations, characterizing a process with stationary laws of evolution, are given by

$$P\left[X(t) = j \mid X(s) = i\right] = P_{ij}(t - s). \tag{6.2.5}$$

for all points $s < t$ in T^+ and states i and j in $\mathfrak{S}$.

From now on, let

$$\boldsymbol{P}(t) = (P_{ij}(t) \mid (i, j) \in \mathfrak{S} \times \mathfrak{S}) \tag{6.2.6}$$

denote a $r \times r$ matrix valued function of $t \in T^+$. Given this matrix valued function, the probability of the event

$$[X(t_k) = i_k \mid k = 0, 1, 2, \ldots, n], \tag{6.2.7}$$

where $t_0 < t_1 < \cdots < t_n$ and $(i_k \mid k = 0, 1, 2, \ldots, n)$ are states in $\mathfrak{S}$ will be assigned the value

$$P\left[X\left(t_k\right) = i_k \mid k = 0, 1, 2, \ldots, n\right] = \prod_{k=1}^{n} P_{i_{k-1}, i_k}\left(t_k - t_{k-1}\right) \qquad (6.2.8)$$

for all integers $n \geq 1$.

The laws of evolution of the process are said to be stationary, if for every $h > 0$ and every $n \geq 1$, the equation is satisfied

$$P\left[X\left(t_k + h\right) = i_k \mid k = 0, 1, 2, \ldots, n\right] = P\left[X\left(t_k\right) = i_k \mid k = 0, 1, 2, \ldots, n\right]. \qquad (6.2.9)$$

It is easy to see that the above assignment of probabilities implies that the process is stationary. This assignment of probabilities also implies that the process has the Markov property, because from the definition, it follows that

$$\begin{aligned} & P\left[X\left(t_n\right) = i_n \mid X\left(t_k\right) = i_k, k = 0, 1, 2, \ldots, n-1\right] \\ &= \frac{P\left[X\left(t_n\right) = i_n \mid k = 0, 1, 2, \ldots, n\right]}{P\left[X\left(t_n\right) = i_k \mid k = 0, 1, 2, \ldots, n-1\right]} \\ &= P\left[X\left(t_n\right) \mid X\left(t_{n-1}\right) = i_{n-1}\right] = P_{i_{n-1} i_n}\left(t_n - t_{n-1}\right). \qquad (6.2.10) \end{aligned}$$

for every $n \geq 1$, time points $t_0 < t_1 < \cdots < t_n$ and states $(i_k \mid k = 0, 1, 2, \ldots, n)$ in $\mathfrak{S}$.

The Markov property also implies that for every $s < t$ in T^+ and pair of states i and j in $\mathfrak{S}$, the equation

$$P_{ij}\left(s + t\right) = \sum_{k \in \mathfrak{S}} P_{ik}\left(s\right) P_{kj}\left(t\right) \qquad (6.2.11)$$

holds. The matrix form of these equations may be expressed as

$$\mathbf{P}\left(s + t\right) = \mathbf{P}\left(s\right) \boldsymbol{P}\left(t\right), \qquad (6.2.12)$$

which is known as the Chapman-Kolmogorov equation. In particular, if $s = 0$, then

$$\boldsymbol{P}\left(t\right) = \boldsymbol{P}\left(0\right) \mathbf{P}\left(t\right) \qquad (6.2.13)$$

for all $t \in T^+$. This equation will be valid for all $t \geq 0$ if

$$\boldsymbol{P}\left(0\right) = \boldsymbol{I}_r, \qquad (6.2.14)$$

where $\boldsymbol{I}_r$ is a $r \times r$ identity matrix.

A problem that arises at this point in the discussion is that of trying to find a method for expressing the matrix valued function $\mathbf{P}(t)$ in some explicit and useful form. To this end, it will be assumed that all the functions in $\mathbf{P}(t)$ are differentiable for all $t \in T^{+}$. By definition, for every $i \in \mathfrak{S}$

$$\frac{d}{ds} P_{ii}(0) = \lim_{h \downarrow 0} \frac{P_{ii}(h) - 1}{h} = q_{ii} \leq 0 \tag{6.2.15}$$

and for every pair of states i and j in $\mathfrak{S}$ such that $i \neq j$

$$\frac{d}{ds} P_{ij}(0) = \lim_{h \downarrow 0} \frac{P_{ij}(h)}{h} = q_{ij} \geq 0. \tag{6.2.16}$$

Because, by assumption,

$$\sum_{j \in \mathfrak{S}} P_{ij}(t) = 1 \tag{6.2.17}$$

for all $t \in T^{+}$ and $i \in \mathfrak{S}$, it follows that

$$\sum_{j \in \mathfrak{S}} \frac{d}{dt} P_{ij}(0) = \sum_{j \in \mathfrak{S}} q_{ij} = 0 \tag{6.2.18}$$

for all $i \in \mathfrak{S}$. Therefore, because $q_{ij} \geq 0$ when $i \neq j$ and $q_{ii} \leq 0$, it also follows that

$$q_{ii} = -\sum_{j \neq i} q_{ij} \tag{6.2.19}$$

for all $i \in \mathfrak{S}$. From now on, for every $i \in \mathfrak{S}$, let

$$q_i = \sum_{j \neq i} q_{ij}, \tag{6.2.20}$$

and define a $r \times r$ matrix $\boldsymbol{Q}$ by

$$\boldsymbol{Q} = \frac{d}{dt} \boldsymbol{P}(0). \tag{6.2.21}$$

Observe that the matrix $\boldsymbol{Q}$ has the property that all off-diagonal elements are non-negative and all diagonal elements are ≤ 0, *i.e.*, non-positive In particular, for the case $r = 2$, this matrix has the form

$$\boldsymbol{Q} = \begin{pmatrix} -q_1 & q_{12} \\ q_{21} & -q_2 \end{pmatrix}, \tag{6.2.22}$$

where $q_1 = q_{12}$ and $q_2 = q_{21}$.

A classical approach to finding an explicit formula for the matrix function $\boldsymbol{P}(t)$ is to find a set of differential equations it satisfies. To this end, let

$$\boldsymbol{P}'(t) = \left(\frac{d}{dt} P_{ij}(t) \right) \tag{6.2.23}$$

represent the matrix of derivatives. Then, fix t in the Chapman-Kolmogorov equation and differentiate with respect to s, which leads to the equation

$$\mathbf{P}^{'}(s+t) = \mathbf{P}^{'}(s)\,\mathbf{P}(t)\,. \tag{6.2.24}$$

Similarly, if t and s are interchanged in the Chapman-Kolmogorov equation and t is fixed, then differentiation with respect to s leads to the equation

$$\mathbf{P}^{'}(t+s) = \mathbf{P}(t)\,\mathbf{P}^{'}(s)\,. \tag{6.2.25}$$

Letting $s = 0$ in these equations, leads to the pair of matrix differential equations

$$\begin{aligned}\mathbf{P}^{'}(t) &= \boldsymbol{Q}\mathbf{P}(t)\\ \mathbf{P}^{'}(t) &= \mathbf{P}(t)\,\boldsymbol{Q},\end{aligned} \tag{6.2.26}$$

which are called the Kolmogorov differential equations. The upper differential equation is known as the backward equation and the lower equation is referred to as the forward equation. The problem of finding an explicit form of the matrix $\mathbf{P}(t)$ will be solved if one can find a function $\mathbf{P}(t)$ which satisfies the initial condition

$$\mathbf{P}(0) = \boldsymbol{I}_r \tag{6.2.27}$$

and the Chapman-Kolmogorov equation

$$\mathbf{P}(s+t) = \mathbf{P}(s)\,\boldsymbol{P}(t)\,. \tag{6.2.28}$$

Furthermore, let $\mathbf{1}_r$ denote a $r \times 1$ column vector with the constant element 1. Then the matrix $\mathbf{P}(t)$ must also satisfy the condition

$$\boldsymbol{P}(t)\,\mathbf{1}_r = \mathbf{1}_r \tag{6.2.29}$$

for all $t \in T^{+}$. In the case of one dimensional functions, the set of differential equations have the form

$$p^{'}(t) = qp(t) = p(t)\,q \tag{6.2.30}$$

with the initial condition $p(0) = 1$ along with the property $p^{'}(s+t) = p^{'}(s)\,p^{'}(t)$. For this simple case, it is easy to see that

$$p(t) = \exp(qt) \tag{6.2.31}$$

is a solution of this system satisfying all the conditions. By analogy with on one dimensional case, one is thus led to consider an exponential matrix function defined by

$$\mathbf{P}(t) = \sum_{k=0}^{\infty} \boldsymbol{Q}^k \frac{t^k}{k!} = \mathbf{I}_r + \boldsymbol{Qt} + \boldsymbol{Q}^2\frac{t^2}{2!} + \cdots\,. \tag{6.2.32}$$

as a solution of the Kolmogorov differential equations satisfying all the stated conditions.

When $\mathbf{P}(t)$ is expressed in the form of an exponential matrix, it is easy to see that this matrix function satisfies the Kolmogorov differential equations as well as the conditions stated above. In this connection note that

$$\mathbf{P}(0) = \mathbf{I}_r. \tag{6.2.33}$$

Next let $\mathbf{0}_r$ denote a $1 \times r$ column vectors with constant value 0. Then the matrix $\boldsymbol{Q}$ satisfies the condition

$$\boldsymbol{Q}\mathbf{1}_r = \mathbf{0}, \tag{6.2.34}$$

which implies that

$$\boldsymbol{Q}^k\mathbf{1}_r = \mathbf{0} \tag{6.2.35}$$

for every integer $k \geq 1$. Hence,

$$\mathbf{P}(t)\,\mathbf{1}_r = \mathbf{I}_r\mathbf{1}_r + \sum_{k=1}^{\infty} \boldsymbol{Q}^k\mathbf{1}_r\frac{t^k}{k!} = \mathbf{1}_r \tag{6.2.36}$$

for all $t \in T^+$. That $\mathbf{P}(t)$ also satisfies the Chapman-Kolmogorov equation follows by noting that

$$\begin{aligned}
\mathbf{P}(s)\,\mathbf{P}(t) &= \left(\sum_{k=0}^{\infty} \boldsymbol{Q}^k\frac{s^k}{k!}\right)\left(\sum_{k=0}^{\infty} \boldsymbol{Q}^k\frac{t^k}{k!}\right) \\
&= \sum_{n=0}^{\infty}\sum_{k=0}^{n} \boldsymbol{Q}^k\frac{s^k}{k!}\boldsymbol{Q}^{n-k}\frac{t^{n-k}}{(n-k)!} \\
&= \sum_{n=0}^{\infty}\frac{1}{n!}\sum_{k=0}^{n} \boldsymbol{Q}^n\binom{n}{k}s^k t^{n-k} \\
&= \sum_{n=0}^{\infty}\frac{\boldsymbol{Q}^n}{n!}(s+t)^n = \mathbf{P}(s+t)\,.
\end{aligned} \tag{6.2.37}$$

These matrix manipulations are justified because the matrix series defining $\mathbf{P}(t)$ converges absolutely for all $t \in T^+$. To see this consider, let $\parallel \cdot \parallel$ denote the norm of a square matrix, Then,

$$\begin{aligned}
\parallel \mathbf{P}(t) \parallel &= \parallel \sum_{k=0}^{\infty} \boldsymbol{Q}^k\frac{t^k}{k!} \parallel \\
&\leq \sum_{k=0}^{\infty} \parallel \boldsymbol{Q} \parallel^k \frac{t^k}{k!} \\
&= \exp\left(\parallel \boldsymbol{Q} \parallel t\right) < \infty
\end{aligned} \tag{6.2.38}$$

for all $t \in T^+$.

From the foregoing discussion, it can be seen that, given a numerical specification of the parameter matrix $\boldsymbol{Q}$ of constants, the laws of evolution of the processes are completely determined. A very useful form of the exponential matrix $\mathbf{P}(t)$ may be expressed in terms of the eigenvalues and eigenvectors of the matrix $\boldsymbol{Q}$. The number λ, real or complex, is said to be an eigenvalue of $\boldsymbol{Q}$ if there exists a $r \times 1$ vector $\boldsymbol{x}$ of numbers such that $\boldsymbol{x} \neq \mathbf{0}_r$ and

$$\boldsymbol{Qx} = \boldsymbol{x\lambda}. \tag{6.2.39}$$

In general, the matrix $\boldsymbol{Q}$ will have r eigenvalues that may be represented as the diagonal matrix

$$\boldsymbol{\Lambda} = \boldsymbol{diag}\left(\lambda_1, \lambda_2, \ldots, \lambda_r\right). \tag{6.2.40}$$

Let the $r \times r$ matrix

$$\boldsymbol{U} = \left(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_r\right) \tag{6.2.41}$$

denote a matrix of the corresponding eigenvectors. That is, the *j-th* column of $\boldsymbol{U}$ is an eigenvector corresponding to the eigenvalue λ_j. Then, after a little reflection it can be seen that

$$\boldsymbol{QU} = \boldsymbol{U\Lambda}. \tag{6.2.42}$$

If the matrix $\boldsymbol{U}$ is non-singular, then

$$\boldsymbol{U}^{-1}\boldsymbol{QU} = \boldsymbol{\Lambda}. \tag{6.2.43}$$

Equivalently,

$$\boldsymbol{Q} = \boldsymbol{U\Lambda U}^{-1} \tag{6.2.44}$$

Therefore,

$$\boldsymbol{Q}^2 = \boldsymbol{U\Lambda U}^{-1}\boldsymbol{U\Lambda U}^{-1} = \boldsymbol{U\Lambda}^2\boldsymbol{U}^{-1} \tag{6.2.45}$$

and by induction, it can be shown that

$$\boldsymbol{Q}^k = \boldsymbol{U\Lambda}^k\boldsymbol{U}^{-1} \tag{6.2.46}$$

for every integer $k \geq 0$.

From this formula, it can be seen that

$$\begin{aligned}\mathbf{P}(t) &= \sum_{k=0}^{\infty} \boldsymbol{Q}^k \frac{t^k}{k!} = \boldsymbol{U}\left(\sum_{k=0}^{\infty} \boldsymbol{\Delta}^k \frac{t^k}{k!}\right)\boldsymbol{U}^{-1} \\ &= \boldsymbol{U}\,\mathbf{exp}\left(\boldsymbol{\Lambda}t\right)\boldsymbol{U}^{-1}.\end{aligned} \tag{6.2.47}$$

But,

$$\mathbf{exp}\left(\mathbf{\Lambda}t\right)=\boldsymbol{diag}\left(\exp\left(\lambda_1 t\right),\exp\left(\lambda_2 t\right),\ldots,\exp\left(\lambda_r t\right)\right). \tag{6.2.48}$$

If an investigator's only resource is pencil and paper, then the derivation of this form of the matrix $\mathbf{P}\left(t\right)$ can be a formidable or even an impossible task, particularly if r is large. Fortunately, however, there is now software packages available with the power to do the necessary computations in either symbolic or numerical form. Indeed, the word processor used to type this manuscript has a symbolic and numerical computation engine attached to it so that, as will be shown in subsequent sections of the chapter, explicit forms of the matrix $\mathbf{P}\left(t\right)$ may be easily derived.

For those readers interested in more background on the exponential matrix function, the books Bellman (1953) and (1960) may be consulted. Further information on this function may be found in an appendix to the book, Capasso (1991).

6.3 Stationary Distributions of Markov Chains in Continuous Time with Stationary Laws of Evolution

As has been demonstrated in the foregoing section, the explicit form of the matrix function $\boldsymbol{P}\left(t\right)$ depends on the eigenvalues and eigenvectors of the matrix $\boldsymbol{Q}$. In this section, attention will be devoted to the case that this matrix is not reducible. That is, there is no permutation of the rows or columns of the $r\times r$ matrix $\boldsymbol{Q}$ such that it may be represented in the partitioned form

$$\boldsymbol{Q}=\begin{pmatrix}\boldsymbol{Q}_{11} & \mathbf{0}\\ \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22}\end{pmatrix}, \tag{6.3.1}$$

where $\boldsymbol{Q}_{11}$ is a $r_1\times r_1$ matrix, $\mathbf{0}$ is a $r_1\times r_2$ matrix with constant element 0, $\boldsymbol{Q}_{21}$ is a $r_2\times r_1$ matrix and $\boldsymbol{Q}_{22}$ is a $r_2\times r_2$ matrix. The numbers r_1 and r_2 are integers ≥ 1 and such that $r_1+r_2=r$. When the matrix $\boldsymbol{Q}$ is not reducible, then all states in the state space $\mathfrak{S}$ communicate in the sense that for every $i\in\mathfrak{S}$ such that $X\left(0\right)=i$, $P_{ij}\left(t\right)>0$ for all $j\in\mathfrak{S}$ and $t>0$. In such chains in continuous time, it seems plausible that the process would eventually converge to a stationary distribution as $t\to\infty$. As will be shown in what follows, there will be a connection between a certain eigenvector of $\boldsymbol{Q}$ and the stationary distribution.

To demonstrate that convergence to a stationary distribution does indeed occur, it will be helpful to view the evolution of the Markov chain from the sample path perspective. In this connection, suppose the initial state of the process at time $t = 0$ is $X(0) = i$ and let T_i denote a random variable representing the duration of stay in state i. After a random stay in state i, the process moves to some state $j \neq i$ with some conditional probability and remains in state j for some random time T_j and then jumps to some state $k \neq j$ and so the process continues. One of the well known properties of stationary Markov jump processes in continuous time is that the Markov property implies that the random variable T_i has an exponential distribution with scale parameter q_i for all $i \in \mathfrak{S}$. In symbols, given that $X(0) = i$, the probability that the process is still in state i at time $t > 0$ is

$$S_i(t) = P[T_i > t] = \exp(-q_i t) \tag{6.3.2}$$

so that the distribution function of T_i is

$$F_i(t) = P[T_i \leq t] = 1 - \exp(-q_i t) \tag{6.3.3}$$

for all $t \in T^+$ and $i \in \mathfrak{S}$. As is well known, the expectation of the random variable T_i in this case is $E[T_i] = 1/q_i$.

Furthermore, it can be shown that, given that the process is in state i and a jump occurs, then the conditional probability that there is a transition to state $j \neq i$ is

$$\pi_{ij} = \frac{q_{ij}}{q_i}. \tag{6.3.4}$$

It will be noted that the properties of a Markov chain in continuous time with stationary laws of evolution just outlined provide a basis for Monte Carlo algorithm for simulating realizations of the process, but no attempt will be made in this section to pursue this observation in more detail.

To demonstrate that the process does indeed converge to a stationary distribution as $t \uparrow \infty$, it will be helpful to have a brief excursion into renewal theory. Let T_{ij} denote a random variable representing the first entrance time into state $j \in \mathfrak{S}$, given that $X(0) = i$, let $F_{ij}(t)$ denote its distribution function and let $f_{ij}(t)$ denote its density function. If $i = j$, then $F_{ii}(t)$ is interpreted as the distribution function of the first return time to i after the first jump from i. Given that $X(0) = i$, the event that the process is in state i at time $t > 0$ may be decomposed into two disjoint events. Either there has been no jump from i during the time interval $(0, t]$ with probability $\exp(-q_i t)$ or there was at least one jump during this time interval and the process returned to i for the first at some time $s < t$ with probability $f_{ii}(s)\,ds.$ and is still in i at time t with probability $P_{ii}(t-s)$. Integrating over all values of s leads to the renewal type integral equation

$$P_{ii}(t) = \exp(-q_i t) + \int_0^t f_{ii}(s)\,P_{ii}(t-s)\,ds. \tag{6.3.5}$$

The behavior of this equation as $t \uparrow \infty$ has been studied by many authors. Let

$$m_{ii} = E[T_{ii}] = \int_0^\infty s f_{ii}(s)\,ds \tag{6.3.6}$$

denote the finite expectation of the random variable T_{ii}. Then, in a fundamental theorem of renewal theory it is stated that

$$\lim_{t\uparrow\infty} P_{ii}(t) = \frac{1}{m_{ii}} \int_0^\infty \exp(-q_i t)\,dt = \frac{1}{q_i m_{ii}}. \tag{6.3.7}$$

By a similar renewal argument, it can be shown that if $i \neq j$, then

$$P_{ij}(t) = \int_0^t f_{ij}(s)\,P_{jj}(t-s)\,ds. \tag{6.3.8}$$

From this equation, it follows that, because

$$\lim_{t\uparrow\infty} \int_0^t f_{ij}(s)\,ds = 1, \tag{6.3.9}$$

the limit of $P_{ij}(t)$ as $t \uparrow \infty$ has the form

$$\lim_{t\uparrow\infty} P_{ij}(t) = \frac{1}{q_j m_{jj}} \tag{6.3.10}$$

for every $j \neq i$. Accounts of renewal theory may be found in the classic books by Feller (1968) and (1966) and also by Karlin and Taylor (1975) and (1981).

Let

$$\pi_j = \frac{1}{q_j m_{jj}} \tag{6.3.11}$$

for all $j \in \mathfrak{S}$ and let

$$\boldsymbol{\pi} \;=\; (\pi_1, \pi_2, \ldots, \pi_r) \tag{6.3.12}$$

denote a $1 \times r$ vector of these symbols. A question that arises at this point in the discussion is whether the elements in the vector $\boldsymbol{\pi}$ have the property

$$\sum_{j=1}^{r} \pi_j = 1, \tag{6.3.13}$$

which would qualify it as a candidate for a stationary distribution. To see that this equation holds, observe that every $i \in \mathfrak{S}$

$$\sum_{j=1}^{r} P_{ij}(t) = 1 \tag{6.3.14}$$

for all $t \in T^{+}$. Therefore,

$$\lim_{t \uparrow \infty} \sum_{j=1}^{r} P_{ij}(t) = \sum_{j=1}^{r} \lim_{t \uparrow \infty} P_{ij}(t) = \sum_{j=1}^{r} \pi_j = 1, \tag{6.3.15}$$

which shows that the vector $\boldsymbol{\pi}$ is a probability distribution and thus qualifies as a candidate for a stationary distribution.

As a first step in a quest for a more precise meaning for the term, stationary distribution, let $\mathbf{1}_r$ denote a $r \times 1$ column vector with constant element 1, and define a $r \times r$ matrix $\boldsymbol{A}$ by the equation

$$\boldsymbol{A} = \mathbf{1}_r \boldsymbol{\pi}. \tag{6.3.16}$$

Note that each row of this matrix is the vector $\boldsymbol{\pi}$. Given this matrix, it follows from the foregoing discussion that

$$\lim_{t \uparrow \infty} \boldsymbol{P}(t) = \boldsymbol{A}. \tag{6.3.17}$$

From this limit, it follows that there is a $r \times r$ matrix $\boldsymbol{R}(t)$ such that

$$\boldsymbol{P}(t) = \boldsymbol{A} + \boldsymbol{R}(t) \tag{6.3.18}$$

for all $t \in T^{+}$, where $\boldsymbol{R}(t) \to \mathbf{0}$, a $r \times r$ zero matrix as $t \uparrow \infty$.

Now recall that the Chapman-Kolmogorov equation has the matrix form

$$\boldsymbol{P}(s+t) = \boldsymbol{P}(s)\boldsymbol{P}(t) \tag{6.3.19}$$

for all s and t in T^{+}. By fixing $t \in T^{+}$ and letting $s \uparrow \infty$ in this equation, it follows that the matrix $\boldsymbol{A}$ satisfies the equation

$$\boldsymbol{A} = \boldsymbol{A}\boldsymbol{P}(t) \tag{6.3.20}$$

for all $t \in T^{+}$, Each row of this matrix is of the form

$$\boldsymbol{\pi} = \boldsymbol{\pi}\boldsymbol{P}(t) \tag{6.3.21}$$

for all $t \in T^{+}$. This equation demonstrates that $\boldsymbol{\pi}$ is the stationary distribution of the process in the sense that if $\boldsymbol{\pi}$ is the initial distribution, then after $t > 0$ time units of evolution, the distribution of the process is still $\boldsymbol{\pi}$ for all $t > 0$. The next problem that arises in applying the theory in genetics is that of finding some method of calculating the vector $\boldsymbol{\pi}$ in either symbolic or numerical forms, given the parameter matrix $\boldsymbol{Q}$.

In order to describe a method for expressing $\boldsymbol{\pi}$ in terms of the matrix $\boldsymbol{Q}$, it will be helpful to exploit some of the properties of this matrix. In the previous section, it was shown that in the matrix $\boldsymbol{Q}$ all off-diagonal elements are non-negative, *i.e.*, if $i \neq j$, then $q_{ij} \geq 0$. Then, because

$$q_i = \sum_{j \neq i} q_{ij} \geq 0 \tag{6.3.22}$$

for all $i \in \mathfrak{S}$ and $q_{ii} = -q_i$. the elements on the principal diagonal of $\boldsymbol{Q}$ are non-positive, *i.e.*, $q_{ii} \leq 0$ for all $i \in \mathfrak{S}$. A matrix with these properties is called a quasi-monotone matrix.

Let λ be an eigenvalue of the matrix $\boldsymbol{Q}$ and let $\boldsymbol{x}$ be a corresponding column eigenvector. Then,

$$\boldsymbol{Qx} = \boldsymbol{x}\lambda. \tag{6.3.23}$$

Such column vectors are called right eigenvectors corresponding to λ. Let $\mathfrak{R}_\lambda$ denote the set, space, of all right eigenvectors $\boldsymbol{x}$ satisfying this equation. Similarly, for every eigenvalue λ of $\boldsymbol{Q}$, there is a non-zero row eigenvector vector $\boldsymbol{y}$ such that

$$\boldsymbol{yQ} = \boldsymbol{y}\lambda, \tag{6.3.24}$$

which is called a left eigenvector corresponding to λ. Let $\mathfrak{L}_\lambda$ denote the set, space, of all left eigenvectors satisfying this equation.

According to a theorem on quasi-monotone matrices, there is a simple real eigenvalue ρ of $\boldsymbol{Q}$ such that if λ is any other eigenvalue of $\boldsymbol{Q}$, then $\operatorname{Re}(\lambda) \leq \rho$, where $\operatorname{Re}(\lambda)$ stands for the real part of λ so that

$$\rho = \max\left(\operatorname{Re}\left(\lambda_i\right) \mid i = 1, 2, \ldots, r\right) \tag{6.3.25}$$

Moreover, if $\boldsymbol{y}$ and $\boldsymbol{x}$ are left and right eigenvectors corresponding to ρ, then all the elements in these vectors are positive. Furthermore, the dimension of each of the spaces $\mathfrak{R}_\rho$ and $\mathfrak{L}_\rho$ is one, *i.e.*, any two vectors in these spaces are multiples of each other. A statement of this theorem may be found in an appendix to the book by Capasso (1991) along with some results on the exponential matrix function when the constant matrix is quasi-monotone.

As above, let $\mathbf{1}_r$ denote a $r \times 1$ column vector with constant elements equal to 1. Then,

$$\boldsymbol{Q}\mathbf{1}_r = \mathbf{0}_r, \tag{6.3.26}$$

where $\mathbf{0}_r$ is a $r \times 1$ with all elements equal to 0. It follows, therefore, that $\lambda = 0$ is an eigenvalue of $\boldsymbol{Q}$. It has also been shown that

$$\lim_{t \uparrow \infty} \boldsymbol{P}(t) = \lim_{t \uparrow \infty} \exp\left(\boldsymbol{Q}t\right) = \boldsymbol{A}, \tag{6.3.27}$$

which implies that for all $\lambda \neq 0$, $\mathrm{Re}\,(\lambda) \leq 0$. For if $\mathrm{Re}\,(\lambda) > 0$, for some $\lambda \neq 0$, then the matrix $\boldsymbol{P}\,(t)$ would be unbounded as $t \uparrow \infty$, which is contrary to convergence of $\boldsymbol{P}\,(t)$ to a matrix $\boldsymbol{A}$ with finite elements as $t \uparrow \infty$. Consequently, for all quasi-monotone matrices $\boldsymbol{Q}$, the simple eigenvalue ρ is 0. Therefore, there exists a $1 \times r$ eigenvector $\boldsymbol{y} = \ (y_1, y_2, \ldots, y_r)$ with positive elements such that

$$\boldsymbol{yQ} = \boldsymbol{0}_r^T, \tag{6.3.28}$$

and these elements may be chosen such that

$$\sum_{i=1}^{r} y_i = 1. \tag{6.3.29}$$

With this background in the theory of the quasi-montone matrix $\boldsymbol{Q}$**,** it is a straight-forward exercise to demonstrate a method for finding the stationary vector $\boldsymbol{\pi}$ as a function of the matrix $\boldsymbol{Q}$. For consider the equation

$$\begin{aligned} \boldsymbol{\pi P}\,(t) &= \boldsymbol{\pi} \left(\sum_{k=0}^{\infty} \boldsymbol{Q}^k \frac{t^k}{k!} \right) \\ &= \boldsymbol{\pi} + \boldsymbol{\pi Q} \left(\sum_{k=1}^{\infty} \boldsymbol{Q}^{k-1} \frac{t^k}{k!} \right). \end{aligned} \tag{6.3.30}$$

Then, the equation

$$\boldsymbol{\pi P}\,(t) = \boldsymbol{\pi} \tag{6.3.31}$$

holds for all $t \in T^+$ if, and only if, $\boldsymbol{\pi Q} = \boldsymbol{0}_r^T$, which implies $\boldsymbol{\pi}$ is a left eigenvector corresponding to the eigenvalue $\rho = 0$. The problem of computing the stationary vector $\boldsymbol{\pi}$ thus reduces to computing a left eigenvector of the matrix $\boldsymbol{Q}$ corresponding to the eigenvalue $\rho = 0$, in either a symbolic or numerical form, and normalizing it so that it is a vector of probabilities.

There is a special case that sometimes arises in genetic applications when it is assumed that the matrix $\boldsymbol{Q}$ is symmetric so that $\boldsymbol{Q}^T = \ \boldsymbol{Q}$. It has been noted that $\boldsymbol{Q1}_r = \boldsymbol{0}_r$. Therefore,

$$(\boldsymbol{Q1}_r)^T = \boldsymbol{1}_r^T \boldsymbol{Q} = \boldsymbol{0}_r^T. \tag{6.3.32}$$

Consequently, $\boldsymbol{1}_r^T$ is a left eigenvector of $\boldsymbol{Q}$ corresponding to the eigenvalue $\rho = 0$. From this result, it follows that for all processes with a symmetric matrix $\boldsymbol{Q}$, the stationary distribution is the uniform distribution

$$\boldsymbol{\pi} = \ \left(\pi_i = \frac{1}{r} \mid i = 1, 2, \ldots, r \right) \tag{6.3.33}$$

on the positive integers $1, 2, \ldots, r$. It is also of interest to note that if $\boldsymbol{Q}$ is symmetric, then so is the matrix $\boldsymbol{P}\,(t)$ for all $t \in T^+$.

6.4 Markov Jump Processes as Models for Base Substitutions in the Molecular Evolution of DNA

A molecule of DNA is made of bases which are classified as purines and pyrimidines. The purine bases are

$$\begin{array}{c}\textbf{Purine Bases}\\ A - \text{Adenine}\\ G - \text{Guanine}\end{array} \tag{6.4.1}$$

and the pyrimidine bases are

$$\begin{array}{c}\textbf{Pyrimidine Bases}\\ C - \text{Cytocine}\\ T - \text{Thymine}\\ U - \text{Uracil (RNA)}\end{array} . \tag{6.4.2}$$

When viewed linearly, a molecule of DNA is made up of pairs of strands such that bases always occur in pairs. These pairs of bases are

$$\begin{array}{c}\textbf{Pair Bonds (DNA)}\\ \text{A} \Leftrightarrow \text{T}\\ \text{G} \Longleftrightarrow \text{C}\end{array} . \tag{6.4.3}$$

Observe that these pairs of bonds are such that a purine is always paired with a pyrimidine. Let

$$\mathfrak{S} = (A, G, C, T) \leftarrow (1, 2, 3, 4) \tag{6.4.4}$$

denote the set of bases making up a molecule of DNA, where the symbol ← means correspondence. In this section, an overview of models for a type of mutation called a base changes or substitutions in a molecule of DNA will be considered. Such mutations are also known a nucleotide substitutions. For more information on DNA at the molecular level, the books Nei and Kumar (2000), Li (1997) as well as Strachan and Read (2004) may be consulted. A nucleotide substitution will be said to have occurred, if during the evolutionary process, one base has been substituted for another at some base site in a single strand to DNA.

For example, the substitution $A \to G$ indicates that purine A and been substituted by G, another purine; whereas the substitution $A \to C$ indicates that the pyrimidine C has been substituted for purine A. During the sequencing of the genome of humans and other species, single nucleotide polymorphisms ($SNP's$) have been observed throughout the genomes of many individuals, which are thought to have arisen during evolution due to the process of nucleotide substitution. In what follows, the process of nucleotide substitution in evolutionary time will be modelled as Markov jump processes in continuous time with state space $\mathfrak{S}$, where the symbols are ordered from left to right as in (6.4.4).

Example 1: The Jukes - Cantor Model

This model was proposed by Jukes and Cantor (1969) and it sometimes referred to as the JC69 model, see for example Yang (2006). From the perspective of Markov jump process described in the foregoing sections, the $\boldsymbol{Q}$-matrix of this model has the form

$$\boldsymbol{Q} = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}, \tag{6.4.5}$$

where $\alpha > 0$ is a constant. Under the assumption that the process is Markovian, the distribution function of a random variable T, representing the duration of stay in any state during the molecular evolution of DNA, is

$$P\left[T \leq t\right] = 1 - \exp\left(-3\alpha t\right) \tag{6.4.6}$$

for all $t \in T^{+}$ and $\alpha > 0$. In this model, the expectation of this random variable is

$$E\left[T\right] = \frac{1}{3\alpha} \tag{6.4.7}$$

for every state in $\mathfrak{S}$.

Therefore, if the time unit is a year, then α must be chosen such that α^{-1} is expressed in years. By way of an illustrative example, if $\alpha = 10^{-4}$, then, according to this model, the expected waiting time for any nucleotide substitution to occur would be

$$E\left[T\right] = \frac{10^{4}}{3} = 3,333.333,333 \tag{6.4.8}$$

years. It should be noted that if it were desired to express this time in generations, then, by having an estimate of the generation time expressed in years, this expectation could be converted to generations. It is essential to note that, in this formulation of the mutation process of nucleotide substitution, attention is restricted to only those mutations that occur in DNA that is copied from the DNA of parents and passed on to their offspring from generation to generation. Consequently, the mutations that are modelled in this formulation are with respect to two kinds of DNA; namely nuclear DNA that may contain various combinations of parental male and female DNA that arise through the process of genetic recombination as well as that DNA that does not undergo genetic recombination.

Examples of the latter kind of DNA is that in the mitochondria that is passed from mothers to daughters in lines of female descent and also for certain DNA on the Y-chromosome that is passed from fathers to sons in lines of male descent from generation to generation. It is this passing of DNA form generation to generation that leads to the concept of DNA molecules having long life spans which greatly exceed that of individuals in populations as they evolve. Consequently, in this view of molecular evolution, the idea of expressing mutation rates in units of time, such as years, is a useful way of thinking about base substitutions from the evolutionary perspective.

As the $\boldsymbol{Q}$-matrix of the JC69 model is symmetric, it follows that the stationary distribution of the process is the uniform distribution

$$\boldsymbol{\pi} = \left(\pi_i = \frac{1}{4} \mid i = 1, 2, 3, 4 \right) \tag{6.4.9}$$

expressed as a 1×4 row vector. It is of interest to note that, because $\pi_i = 3\alpha / m_{ii}$, where m_{ii} is the expected return time for nucleotide i, given there was substitution to i at some time during the evolutionary past, the probability $\pi_i = 1/4$ may be used to derive a formula for m_{ii} by using the equation

$$\frac{1}{3\alpha m_{ii}} = \frac{1}{4}. \tag{6.4.10}$$

Solving this equation for m_{ii} leads to the formula

$$m_{ii} = \frac{4}{3\alpha}, \tag{6.4.11}$$

which is valid for all $i \in \mathfrak{S}$. If, for example, $\alpha = 10^{-4}$. then

$$m_{ii} = \frac{4 \times 10^4}{3} = 13,333.333,\ 333,\ 333 \tag{6.4.12}$$

years. Observe that for this assigned value of α, the mean return time for any state $i \in \mathfrak{S}$ is greater than $E\left[T\right]$, the expected waiting time in any state.

Thanks to the symbolic computation engine linked to the word processor used to type this manuscript, it is easy to find the eigenvalues of the matrix

$\boldsymbol{Q}$, which are 0 and -4α. Note that the eigenvalue -4α has multiplicity 3. Let $\mathbf{1}_4$ denote a 4×1 column vector with constant element and let

$$\boldsymbol{A} = \mathbf{1}_4 \pi \tag{6.4.13}$$

denote a $4{\times}4$ matrix such that each row is the vector $\boldsymbol{\pi}$. Then, by calling the symbolic computation engine to compute the exponential matrix $\exp(\boldsymbol{Q}t)$, it is possible to express the result in the form

$$\boldsymbol{P}(t) = \boldsymbol{A} + \boldsymbol{R}(t), \tag{6.4.14}$$

where

$$\boldsymbol{R}(t) = \begin{pmatrix} \frac{3}{4}e^{-4t\alpha} & -\frac{1}{4}e^{-4t\alpha} & -\frac{1}{4}e^{-4t\alpha} & -\frac{1}{4}e^{-4t\alpha} \\ -\frac{1}{4}e^{-4t\alpha} & \frac{3}{4}e^{-4t\alpha} & -\frac{1}{4}e^{-4t\alpha} & -\frac{1}{4}e^{-4t\alpha} \\ -\frac{1}{4}e^{-4t\alpha} & -\frac{1}{4}e^{-4t\alpha} & \frac{3}{4}e^{-4t\alpha} & -\frac{1}{4}e^{-4t\alpha} \\ -\frac{1}{4}e^{-4t\alpha} & -\frac{1}{4}e^{-4t\alpha} & -\frac{1}{4}e^{-4t\alpha} & \frac{3}{4}e^{-4t\alpha} \end{pmatrix}. \tag{6.4.15}$$

Interestingly, the norm of this matrix is

$$\parallel \boldsymbol{R}(t) \parallel = \frac{3}{2}e^{-4t\alpha} \tag{6.4.16}$$

for all $t \in T^+$ This formula provides an interesting and useful way of finding an estimate of the time taken for the process of nucleotide substitution process to converge to the stationary distribution. For consider the equation

$$\parallel \boldsymbol{P}(t) - \boldsymbol{A} \parallel = \parallel \boldsymbol{R}(t) \parallel = \frac{3}{2}e^{-4t\alpha}, \tag{6.4.17}$$

and suppose a t is chosen such that

$$\frac{3}{2}e^{-4t\alpha} = 10^{-6}. \tag{6.4.18}$$

Then

$$t = -\frac{1}{4\alpha} \times \ln\left(\frac{2}{3} \times 10^{-6}\right) = 35,552,439,165,\ 181 \tag{6.4.19}$$

years if $\alpha = 10^{-4}$.

In summary, imagine watching the evolution on one base site of a molecule of DNA as it is passed from generation to generation over a long period of years. Then, if $\alpha = 10^{-4}$ and the site is occupied by base $i \in \mathfrak{S}$ at time $t = 0$, then the waiting time to the first jump, nucleotide substitution, would on average be about 3,333 years. Following this first jump, it would take about 13,333 years on average to return to nucleotide i. Finally, after about 35,552 years of evolution, the probability of seeing any of the four bases at this particular site would be 1/4. Furthermore, in a large sample of individuals from a population that had evolved for about 35,552 years, the frequency of each base at this particular site of the DNA molecule would be 1/4.

Example 2: Symmetric Kimura Model

In this model, which was introduced by Kimura (1980), the $\boldsymbol{Q}$-matrix of a nucleotide substitution process has the form

$$Q = \begin{pmatrix} -\alpha-2\beta & \alpha & \beta & \beta \\ \alpha & -\alpha-2\beta & \beta & \beta \\ \beta & \beta & -\alpha-2\beta & \alpha \\ \beta & \beta & \alpha & -\alpha-2\beta \end{pmatrix}, \tag{6.4.20}$$

where the rate parameters α and β are positive. This model is often referred to as K80. In this formulation, it is assumed that there is a single rate α for base changes within both the purine and pyrimidine classes of nucleotides, which are referred to as transitions. Changes form purines to pyrimidines and from pyrimidines to purines and referred to as transversions and it is assumed that such changes are governed by a single rate parameter β.

Like the Jukes-Cantor model, this matrix is symmetric so that the stationary distribution of the process is again the 1×4 vector $\boldsymbol{\pi}$ such that $\pi_i = 1/4$ for all $i = 1,2,3,4$. From the form of the matrix $\boldsymbol{Q}$, it also follows that the expected sojourn time in any state $i \in \mathfrak{S}$ is

$$E\left[T\right] = \frac{1}{\alpha+2\beta}. \tag{6.4.21}$$

As is easily shown by calling the symbolic computation engine, the eigenvalues of the matrix $\boldsymbol{Q}$ are $0, -4\beta$ and $-2\alpha-2\beta$, and the matrix $\boldsymbol{R}(t) = (r_{ij}(t))$ in the representation $\boldsymbol{P}(t) = \boldsymbol{A}+\ \boldsymbol{R}(t)$ has the partitioned form

$$\boldsymbol{R}(t) = \begin{pmatrix} \boldsymbol{R}_{11}(t) & \boldsymbol{R}_{12}(t) \\ \boldsymbol{R}_{21}(t) & \boldsymbol{R}_{22}(t) \end{pmatrix}, \tag{6.4.22}$$

where

$$\boldsymbol{R}_{11}(t) = \begin{pmatrix} \frac{1}{4}e^{-4t\beta} + \frac{1}{2}e^{-2t\alpha-2t\beta} & \frac{1}{4}e^{-4t\beta} - \frac{1}{2}e^{-2t\alpha-2t\beta} \\ \frac{1}{4}e^{-4t\beta} - \frac{1}{2}e^{-2t\alpha-2t\beta} & \frac{1}{4}e^{-4t\beta} + \frac{1}{2}e^{-2t\alpha-2t\beta} \end{pmatrix} \tag{6.4.23}$$

and

$$\boldsymbol{R}_{12}(t) = \begin{pmatrix} -\frac{1}{4}e^{-4t\beta} & -\frac{1}{4}e^{-4t\beta} \\ -\frac{1}{4}e^{-4t\beta} & -\frac{1}{4}e^{-4t\beta} \end{pmatrix}. \tag{6.4.24}$$

As the form of the matrix $\boldsymbol{Q}$ matrix $\boldsymbol{R}(t)$ is symmetric for all $t \in T^{+}$, the conditions

$$\begin{aligned} \boldsymbol{R}_{11}(t) &= \boldsymbol{R}_{22}(t) \\ \boldsymbol{R}_{12}(t) &= \boldsymbol{R}_{21}(t) \end{aligned} \tag{6.4.25}$$

are also satisfied.

To find a useful bound for the norm of the matrix $\boldsymbol{R}(t)$, it is interesting to observe that

$$\sum_{j=1}^{4} \mid r_{ij}(t) \mid \leq e^{-4t\beta} + e^{-2t\alpha - 2t\beta} \tag{6.4.26}$$

for every $i \in \mathfrak{S}$ and all $t \in T^{+}$. Therefore, the norm of $\boldsymbol{R}(t)$ satisfies the inequality

$$\parallel \boldsymbol{R}(t) \parallel \leq e^{-4t\beta} + e^{-2t\alpha - 2t\beta} \tag{6.4.27}$$

for all $t \in T^{+}$. Thus, in principle, with the help of a root finding method, one could find a t such that

$$\parallel \boldsymbol{P}(t) - \boldsymbol{A} \parallel \leq \ e^{-4t\beta} + e^{-2t\alpha - 2t\beta} = 10^{-6} \tag{6.4.28}$$

as an estimate of the time taken for the process to converge to the stationary distribution, given any values of the parameters α and β.

Example 3: Generalizations of the Kimura Model

One generalization of the Kimura model is that when it is assumed that the substitution rates from purines to pyrimidines and from pyrimidines to purines differ, which gives rise to a rate matrix of the form

$$\boldsymbol{Q} = \begin{pmatrix} -\alpha - 2\gamma & \alpha & \gamma & \gamma \\ \alpha & -\alpha - 2\gamma & \gamma & \gamma \\ \delta & \delta & -\alpha - 2\delta & \alpha \\ \delta & \delta & \alpha & -\alpha - 2\delta \end{pmatrix}. \tag{6.4.29}$$

According to this model, if the process is in a purine state and time $t = 0$, then the expected waiting time to the next jump is

$$E[T_i] = \frac{1}{\alpha + 2\beta} \tag{6.4.30}$$

for $i - 1, 2$. But, if at time $t = 0$ the process is in a pyrimidine state, then this expected waiting time is

$$E[T_i] = \frac{1}{\alpha + 2\delta} \tag{6.4.31}$$

for $i = 3, 4$. As can be seen by inspection, the matrix $\boldsymbol{Q}$ is not symmetric in this case.

With this generalization of the rate matrix $\boldsymbol{Q}$, the symbolic computation engine was capable of finding useful expressions for its eigenvalues, which are 0, $2\alpha - 2\gamma, -2\alpha - 2\delta$ and $-2\gamma - 2\delta$. Each of these eigenvalues had multiplicity one. The engine also yielded the result

$$\left(\tfrac{1}{\gamma}\delta \; \tfrac{1}{\gamma}\delta \; 1 \; 1 \right) \tag{6.4.32}$$

as a left eigenvector corresponding to the eigenvalue $\rho = 0$. If this vector is normalized, then one obtains the vector

$$\boldsymbol{\pi} = \frac{1}{2(\gamma + \delta)} \left(\delta \; \delta \; \gamma \; \gamma \right), \tag{6.4.33}$$

which is an symbolic representation of the stationary distribution of the process, and is valid for all values of α, γ and δ.

From this formula for the stationary distribution, it is also possible to derive formulas for the m_{ii}, the expectation of the time taken for the process to return to state i for the first time, given it was in state i at some previous time. For the case of purine states, the equation

$$\frac{1}{(\alpha + 2\delta)\, m_{ii}} = \frac{\delta}{2(\gamma + \delta)} \tag{6.4.34}$$

is valid. Therefore,

$$m_{ii} = \frac{2(\gamma + \delta)}{(\alpha + 2\delta)\,\delta} \tag{6.4.35}$$

for $i = 1, 2$. Given numerical values of the parameters α, γ and δ, a numerical value of m_{ii} could be calculated. However, for this model, the symbolic computation engine did not yield a useful symbolic formula for the matrix $\boldsymbol{P}(t)$ that could be described and printed on one page. Therefore, it would be very difficult to describe a useful procedure for estimating the time taken for the process to converge to the stationary distribution. However, it was possible to obtain useful representations for the matrix $\boldsymbol{P}(t)$ when rational values of the parameters of the parameters α, γ and δ were specified.

Another generalization of the Kimura model is that due to Blaisdell (1985). For this model, the rate matrix has the form

$$\boldsymbol{Q} = \begin{pmatrix} -\alpha - 2\gamma & \alpha & \gamma & \gamma \\ \beta & -\beta - 2\gamma & \gamma & \gamma \\ \delta & \delta & -\beta - 2\delta & \beta \\ \delta & \delta & \alpha & \alpha - 2\delta \end{pmatrix}, \tag{6.4.36}$$

where α, β, γ and δ are positive parameters. Formulas for the stationary distribution of this process have been published, but these formulas will not be reproduced here. As it turned out, the symbolic computation engine did not, in this case, yield a useful formula for the left eigenvector corresponding to the eigenvalue $\rho = 0$. To find this distribution, an interested reader may wish to try to solve the equation

$$\boldsymbol{\pi Q} = \mathbf{0} \tag{6.4.37}$$

symbolically, where $\mathbf{0}$ is a 1×4 vectors of zeroes.

It was possible, however, to specify numerical values of the parameters and use the exponential matrix to find a numerical form of the stationary distribution. For example, when the parameters were assigned the values,

$$\alpha = 10^{-4}$$

$$\beta = 2 \times 10^{-4}$$

$$\gamma = 3 \times 10^{-4}$$

$$\delta = 4 \times 10^{-4},$$

and when the elements of the matrix $\boldsymbol{Q}$ were expressed in decimal form, then one could use the computation engine to define the matrix

$$\boldsymbol{P}(t) = \exp(\boldsymbol{Q} \times \boldsymbol{t})$$

for all $t \in T^{+}$. According to the theory developed in the foregoing sections, if this matrix can be computed for a large value of t, then the function $\boldsymbol{P}(t)$ will be matrix with numerical elements such that each row is the stationary distribution of the process for these assigned values of the parameters. To illustrate this method, if t is assigned the value $t = 10^9$, then the resulting matrix $\boldsymbol{P}\left(10^9\right)$, when truncated to six places, is

$$\begin{pmatrix} 0.317\,460 & 0.253\,968 & 0.194\,805 & 0.233\,766 \\ 0.317\,460 & 0.253\,968 & 0.194\,805 & 0.233\,766 \\ 0.317\,460 & 0.253\,968 & 0.194\,805 & 0.233\,766 \\ 0.317\,460 & 0.253\,968 & 0.194\,805 & 0.233\,766 \end{pmatrix}. \tag{6.4.38}$$

It is of interest to observe for these parameter values that all the elements of the stationary vector are different and in a population that is in equilibrium with respect to a particular site on a DNA molecule, over half of the individuals would exhibit a purine at this site and somewhat less than half would exhibit pyrimidines. The procedure just described could also be used, through an approximation process, to estimate the time taken for the process to converge to a stationary distribution for these assigned parameter values.

Countless models could be described that would generalize the Kimura model. For example, the most general symmetric rate matrix would have the six parameter form

$$\boldsymbol{Q} = \begin{pmatrix} -(\alpha+\beta+\gamma) & \alpha & \beta & \gamma \\ \alpha & -(\alpha+\delta+\epsilon) & \delta & \epsilon \\ \beta & \delta & -(\beta+\delta+\zeta) & \zeta \\ \gamma & \epsilon & \zeta & -(\gamma+\epsilon+\zeta) \end{pmatrix}, \tag{6.4.39}$$

where all the rate parameters are positive.

6.5 Processes with Preassigned Stationary Distributions

Other approaches to constructing rate matrices for nucleotide substitutions at a particular site is that of preassigning a stationary distribution of the process. To this end, let the 1×4 vector

$$\boldsymbol{\pi} = \ (\pi_i \mid i \in \mathfrak{S}) \tag{6.5.1}$$

denote some preassigned stationary distribution such that $\pi_i > 0$ for all i and

$$\sum_{i=1}^{4} \pi_i = 1. \tag{6.5.2}$$

Then, consider the problem of constructing a 4×4 rate matrix $\boldsymbol{Q}$ such that $\boldsymbol{\pi}$ is the stationary distribution. Evidently, investigators have been motivated to construct such matrices as not only as part of a quest for more realism in formulations but also by the desire to use some empirically observed frequencies of nucleotides in a sample of sequenced DNA as a stationary distribution.

Among the first to find a solution to this problem was Felsenstein (1981), who introduced the rate matrix

$$\boldsymbol{Q} = \begin{pmatrix} -\upsilon(1-\pi_1) & \upsilon\pi_2 & \upsilon\pi_3 & \upsilon\pi_4 \\ \upsilon\pi_1 & -\upsilon(1-\pi_2) & \upsilon\pi_3 & \upsilon\pi_4 \\ \upsilon\pi_1 & \upsilon\pi_2 & -\upsilon(1-\pi_3) & \upsilon\pi_4 \\ \upsilon\pi_1 & \upsilon\pi_2 & \upsilon\pi_3 & -\upsilon(1-\pi_4) \end{pmatrix}, \tag{6.5.3}$$

where $\upsilon > 0$ is a parameter. It is easy to see that $\boldsymbol{\pi}$ is indeed that stationary vector for this matrix. For, consider multiplying $\boldsymbol{\pi}$ by the first column of this matrix, which results in the expression

$$\upsilon\pi_1(-(1-\pi_1)+\pi_2+\pi_3+\pi_4) = 0. \tag{6.5.4}$$

By similar arguments, it can be shown that

$$\boldsymbol{\pi Q} = \mathbf{0}, \tag{6.5.5}$$

where $\mathbf{0}$ is a 1×4 vector of zeros. Therefore, $\boldsymbol{\pi}$ is the stationary vector for the matrix $\boldsymbol{Q}$. It is also interesting to note that the expected sojourn time in state $i \in \mathfrak{S}$ is

$$E\left[T_i\right] = \frac{1}{\upsilon\left(1 - \pi_i\right)} \tag{6.5.6}$$

so that the magnitude of this expectation is large when the parameter υ is small. This model has been labeled $F81$ in the literature.

Subsequently, this matrix was modified, Felsenstein and Churchill (1996), to obtain a rate matrix that may be represented in the partitioned form

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} \\ \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22} \end{pmatrix}. \tag{6.5.7}$$

The 2×2 matrix $\boldsymbol{Q}_{11}$ has the form

$$\boldsymbol{Q}_{11} = \begin{pmatrix} -\upsilon((1 - \pi_1) - f_{12}) & \upsilon\left(\pi_2 + f_{12}\right) \\ \upsilon\left(\pi_1 + f_{21}\right) & -\upsilon((1 - \pi_2) - f_{21}) \end{pmatrix}, \tag{6.5.8}$$

where f_{12} and f_{21} are defined by the equations

$$\begin{aligned} f_{12} &= \frac{\kappa\pi_2}{\pi_1 + \pi_2} \\ f_{21} &= \frac{\kappa\pi_1}{\pi_1 + \pi_2} \end{aligned} \tag{6.5.9}$$

and $\upsilon > 0$ and $\kappa > 0$ are parameters. Similarly, if f_{34} and f_{43} are defined by the equations,

$$\begin{aligned} f_{34} &= \frac{\kappa\pi_4}{\pi_3 + \pi_4} \\ f_{43} &= \frac{\kappa\pi_3}{\pi_3 + \pi_4} \end{aligned} \tag{6.5.10}$$

then the 2×2 matrix $\boldsymbol{Q}_{22}$ has the form

$$\boldsymbol{Q}_{22} = \begin{pmatrix} -\upsilon((1 - \pi_3) - f_{34}) & \upsilon\left(\pi_4 + f_{34}\right) \\ \upsilon\left(\pi_3 + f_{43}\right) & -\upsilon((1 - \pi_4) - f_{43}) \end{pmatrix}. \tag{6.5.11}$$

To complete the definition of the rate matrix for this model, the off-diagonal matrices $\boldsymbol{Q}_{12}$ and $\boldsymbol{Q}_{21}$ have the forms

$$\boldsymbol{Q}_{12} = \begin{pmatrix} \upsilon\pi_3 & \upsilon\pi_4 \\ \upsilon\pi_3 & \upsilon\pi_4 \end{pmatrix} \tag{6.5.12}$$

and

$$\boldsymbol{Q}_{21} = \begin{pmatrix} \upsilon\pi_1 & \upsilon\pi_2 \\ \upsilon\pi_1 & \upsilon\pi_2 \end{pmatrix}. \tag{6.5.13}$$

For this formulation, a key observation in proving that $\boldsymbol{\pi}$ is the stationary vector of the matrix $\boldsymbol{Q}$ is in the expression

$$-\pi_1 f_{12} + \pi_2 f_{21} = \frac{\pi_1 \pi_2}{\pi_1 + \pi_2} (\kappa - \kappa) = 0. \tag{6.5.14}$$

A similar equation holds for the constants f_{34} and f_{43}. For this model, the expected waiting time in state $i = 1$ is

$$E[T_1] = \frac{1}{\upsilon((1-\pi_1) - f_{12})}. \tag{6.5.15}$$

Because this number must be positive, it follows that the parameter κ must be chosen such that

$$((1-\pi_1) - f_{12}) > 0. \tag{6.5.16}$$

Other formulas could be written down for the expectations $E[T_i]$ for $i = 2, 3, 4$, and from these formulas it follows that κ must be chosen such that all these expectations are also positive.

Hasegawa et al. (1985) also introduced a model with some vector $\boldsymbol{\pi}$ as the preassigned stationary distribution, which is different from those just described. In the literature this formulation is referred as the HKY model. For this model, the rate matrix is chosen as

$$\boldsymbol{Q} = \begin{pmatrix} -(\upsilon\pi_2 + \nu\theta_1) & \upsilon\pi_2 & \nu\pi_3 & \nu\pi_4 \\ \upsilon\pi_1 & -(\upsilon\pi_1 + \nu\theta_1) & \nu\pi_3 & \nu\pi_4 \\ \nu\pi_1 & \nu\pi_2 & -(\nu\theta_2 + \upsilon\pi_4) & \upsilon\pi_4 \\ \nu\pi_1 & \nu\pi_2 & \upsilon\pi_3 & -(\nu\theta_2 + \upsilon\pi_3) \end{pmatrix}, \tag{6.5.17}$$

where $\theta_1 = \pi_3 + \pi_4$, $\theta_2 = \pi_1 + \pi_2$ and ν and υ are positive parameters. In this case, the expected sojourn time in state $i = 1$ is

$$E[T_1] = \frac{1}{\upsilon\pi_2 + \nu\theta_1}, \tag{6.5.18}$$

which is positive for all parameter values. It can also be seen that for this model, all expected sojourn times in states would will be positive for all positive parameter values. The rate matrix for this model also has the interesting property that it is possible to express all eigenvalues as well as left and right eigenvectors in simple symbolic forms, and, for those readers who are interested in these formulas the original paper or the book Ewens and Grant (2005) may be consulted.

There is also a more general form of a rate matrix with a preassigned stationary distribution that was given by Tavaré (1986). For this model, the rate matrix has the symbolic form

$$\boldsymbol{Q} = \begin{pmatrix} -\upsilon\eta_1 & \upsilon\eta_{12}\pi_2 & \upsilon\eta_{13}\pi_3 & \upsilon\eta_{14}\pi_4 \\ \upsilon\eta_{21}\pi_1 & -\upsilon\eta_2 & \upsilon\eta_{23}\pi_3 & \upsilon\eta_{24}\pi_4 \\ \upsilon\eta_{31}\pi_1 & \upsilon\eta_{32}\pi_2 & -\upsilon\eta_3 & \upsilon\eta_{34}\pi_4 \\ \upsilon\eta_{41}\pi_1 & \upsilon\eta_{42}\pi_2 & \upsilon\eta_{43}\pi_3 & -\upsilon\eta_4 \end{pmatrix}, \tag{6.5.19}$$

where

$$\eta_i = \sum_{j \neq i} \eta_{ij}\pi_j \tag{6.5.20}$$

for $i = 1, 2, 3, 4$. For $\boldsymbol{\pi}$ times the first column, the condition

$$\pi_1 \upsilon \left(-\eta_1 + \sum_{j \neq 1} \pi_j \eta_{j1} \right) = 0 \tag{6.5.21}$$

arises, and holds, if, and only if,

$$\eta_1 = \sum_{j \neq 1} \pi_j \eta_{j1}. \tag{6.5.22}$$

In general, $\boldsymbol{\pi}$ is the stationary distribution of the process if, and only if, the condition

$$\eta_i = \sum_{j \neq i} \pi_j \eta_{ji} = \sum_{j \neq i} \eta_{ij}\pi_j \tag{6.5.23}$$

holds for every i. Therefore, in choosing the η-parameters, care must be exercised to ensure that all these conditions are satisfied. It will be noted that the three examples described above are all special cases of this more general model. Choosing parameters for this version of a rate matrix is not a straight-forward problem; consequently methods for constructing matrices in this class will not be pursued further here. In the next section, a numerical example of a rate matrix will be given in which 12 parameters may be chosen freely.

6.6 A Numerical Example for a Class of Twelve Parameters

In this section, rather than confining attention to a particular nucleotide site, it will be supposed that a set of n sites on a DNA molecule are under consideration and a single rate matrix will be used to model nucleotide substitutions at all sites under consideration. When constructing a general

4×4 rate matrix $\boldsymbol{Q}$ of such nucleotide substitutions, we are free to choose 12 parameters. A numerical example of such a matrix is

$$\boldsymbol{Q} = \begin{pmatrix} -q_1 & 10^{-4.0} & 10^{-5.1} & 10^{-5.111} \\ 10^{-4.12} & -q_2 & 10^{-5.13} & 10^{-5.14} \\ 10^{-5.131} & 10^{-5.32} & -q_3 & 10^{-4.122} \\ 10^{-5.123} & 10^{-5.124} & 10^{-4.13} & -q_4 \end{pmatrix}, \tag{6.6.1}$$

where q_1 is defined by

$$q_1 = 10^{-4.0} + 10^{-5.1} + 10^{-5.111} \tag{6.6.2}$$

and the other $q's$ are defined similarly. Observe that this is an example of a matrix with 12 different rate parameters and suppose time is expressed in years. Then, for this matrix, the expected sojourn time in state 1 is

$$E\left[T_1\right] = \frac{1}{q_1} = 8,643.946,\ 317,\ 535 \tag{6.6.3}$$

years. Of course similar numbers could be calculated for the expected sojourn times in the other states.

A question that arises at this point is whether it would be plausible to use these properties of expected sojourn times in states to construct a plausible parameter space for the 12 parameters that determine a rate matrix, given some prior information or ideas about an evolutionary process of nucleotide substitution under consideration. For example, suppose an investigator has two samples of DNA from two species who were thought to have descended from a common ancestor who lived t_y years ago. Then, to help fix ideas, if $t = 0$ is viewed as the time the two species diverged, then the interval of time since the two species diverged is $(0, t_y)$.

Also suppose that these samples have been aligned so that it was possible to get a count of the number n_b of base sites out of a total of n sites at which the two samples displayed different nucleotides. Then, by definition, n_b/t_y would be a rough estimate of the rate of mutation, base substitution, per unit time during the evolutionary time interval $(0, t_y)$. Such an estimate would provide some ideas as to the magnitude of the rate parameters in a matrix $\boldsymbol{Q}$. In the above example, it was assumed that these rates were roughly in the interval $(10^{-5.1234}, 10^{-4})$, which was viewed as a plausible parameter space.

From the perspective of a Markov jump process in continuous time, the number n_b would also provide some information on the number of jumps that may have occurred in the time interval $(0, t_y)$ since the two species diverged. Evidently, this number would be a minimal estimate of the number of jumps that occurred in this time interval, because returns to a base at each site would not have been counted. Let m_{st}. denote that average sojourn time in a state, then roughly $n_b m_{st} = t_y$ so that $m_{st} = t_y/n_y$ would provide a preliminary estimate of the magnitude of sojourn times in each state. By using such rough estimates one could obtain some idea as to plausibility of the assigned rates in a matrix $\boldsymbol{Q}$.

Given the rate matrix $\boldsymbol{Q}$, other calculations could be made to assess the implications of the chosen rates. For example, it is feasible to write a computer program to compute the matrix $\boldsymbol{P}(t) = \exp(\boldsymbol{Q} \times t)$ as a function of $t > 0$. When the matrix function $\boldsymbol{P}(t)$ is evaluated at some large number, for example $t = 10^9$, then the computer would return a 4×4 matrix of numerical values. In particular, for the rate matrix listed above the computer returned the matrix

$$\boldsymbol{A} = \begin{pmatrix} 0.208\,202 & 0.265\,757 & 0.263\,373 & 0.262\,666 \\ 0.208\,202 & 0.265\,757 & 0.263\,373 & 0.262\,666 \\ 0.208\,202 & 0.265\,757 & 0.263\,373 & 0.262\,666 \\ 0.208\,202 & 0.265\,757 & 0.263\,373 & 0.262\,666 \end{pmatrix}, \qquad (6.6.4)$$

which has been truncated to six decimal places. Observe that each row of this matrix is an estimate of the stationary distribution of a process with rate matrix $\boldsymbol{Q}$. This result could also have been obtained by finding the left eigenvector of $\boldsymbol{Q}$ corresponding to the eigenvalue $\rho = 0$.

From this result, one could also get an estimate of the time taken for the process to converge to the stationary distribution. For consider the matrix

$$\boldsymbol{R}(t) = \boldsymbol{P}(t) - \boldsymbol{A} \qquad (6.6.5)$$

and the equation

$$\| \boldsymbol{R}(t) \| = \epsilon, \qquad (6.6.6)$$

where ϵ is some small assigned number. Then, when it is possible to compute values of the matrix $\boldsymbol{R}(t)$ rapidly, it would be possible to find a t_s, the time taken to converge to the stationary distribution, given the assigned rates in the matrix $\boldsymbol{Q}$.

6.7 Falsifiable Predictions of Markov Models of Nucleotide Substitutions

A property of all Markov models of nucleotide substitutions considered so far is that as t becomes large the matrix $\boldsymbol{P}(t)$ converges to the limit

$$\lim_{t\uparrow\infty} \boldsymbol{P}(t) = \begin{pmatrix} \pi_1 & \pi_2 & \pi_3 & \pi_4 \\ \pi_1 & \pi_2 & \pi_3 & \pi_4 \\ \pi_1 & \pi_2 & \pi_3 & \pi_4 \\ \pi_1 & \pi_2 & \pi_3 & \pi_4 \end{pmatrix}, \tag{6.7.1}$$

which implies that the evolutionary process has reached a state of statistical equilibrium such that the nucleotides in a set of n sites are distributed independently. In other words, a set of n sites may be viewed as an independent and identically distributed sample from the stationary distribution $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3, \pi_4)$. This mathematical result may also be interpreted as one of the predictions of the model. In principle, therefore, it is reasonable to conduct some statistical tests for independence to provide evidence as to whether a proposed Markov model is a good predictor of what was actually observed in a sequence of n nucleotides. If the observed sequence fails to pass such tests, then this evidence could be interpreted as an empirical falsification of a proposed Markov model of nucleotide substitutions. This point of view would be consistent with the ideas of falsifiabilty of scientific hypotheses developed in the book, Popper (1939), on the logic of scientific discovery.

A question that naturally arises as this point in the discussion is that of a choice of statistical tests for independence among the many tests that could be chosen. By way of an illustration, suppose an investigator partitions a set of n sites into $k \geq 1$ subsets such that

$$\sum_{i=1}^{k} n_i = n, \tag{6.7.2}$$

where n_i is the number of sites in the *i-th* subset. For every i let n_{ij} denote the observed number nucleotides of type $j = 1, 2, 3, 4$ among the n_i nucleotides in the *i-th* subset. Such that

$$\sum_{j=1}^{4} n_{ij} = n_i \tag{6.7.3}$$

for every i. Then suppose the data are arranged in a $k \times 4$ array of the form

$$\begin{pmatrix} n_{11} & n_{12} & n_{13} & n_{14} \\ n_{21} & n_{22} & n_{23} & n_{24} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ n_{k1} & n_{k2} & n_{k3} & n_{k4} \end{pmatrix}. \tag{6.7.4}$$

Given this array of data, one could test the hypothesis that the rows and columns of this array are independent as would be expected under the hypothesis that the evolutionary process is in a state governed by stationary distribution of the Markov process. One could also use this data to estimate the probabilities in the stationary distribution $\boldsymbol{\pi}$. For example, an estimate of the probability π_j would be

$$\widehat{\pi}_j = \frac{1}{n} \sum_{i=1}^{k} n_{ij} \tag{6.7.5}$$

for $j = 1, 2, 3, 4$. An investigator could then use these estimates to test the hypotheses that the rows of the array are independent samples from the stationary distribution $\boldsymbol{\pi}$ as one would expect if the evolutionary process were in a state of statistical equilibrium.

If the data passes tests for independence of the type just discussed, then it would reasonable to suppose that the evolutionary processes underlying the generation of the data seemed to resemble data generated by Markov nucleotide substitution process in statistical equilibrium. If, however, the data failed to pass such tests for independence, then at least two interpretations may seem plausible. One interpretation could be stated as a hypothesis that the evolutionary process had not yet reached a state of statistical equilibrium. In this connection an estimate of the time taken by the process to converge to an equilibrium would be helpful and informative in making an assessment as to whether this hypothesis were plausible. A second interpretation may consist of the idea that a four-state Markov process of nucleotide substitution was not an adequate framework for understanding the data and that it was necessary to consider more general formulations that would accommodate stationary distributions that would allow for the possibility that a sequence of nucleotides were not independent and identically distributed observations from some stationary distribution. Such possibilities will be explored in subsequent sections.

6.8 Position Dependent Nucleotide Substitution Models

Position dependent models of nucleotide substitutions have been formulated to accommodate the idea that rates of nucleotide substitutions at a given site depend on the nucleotide at an adjacent site. For example, consider a nucleotide sequence with n sites and suppose n is an even number. Then, such a sequence may be viewed as a sequence of pairs of nucleotides, which are referred to as dinucleotides. As a first step toward developing a Markov model with rate matrix $\boldsymbol{Q}$ for nucleotide substitutions that may occur for the case of dinucleotides, it will be helpful to enumerate the set of possible pairs that may be generated from the four bases

$$(A, G, C, T) \leftarrow (1, 2, 3, 4) \tag{6.8.1}$$

making up a DNA molecule.

The set $\mathfrak{P}$ of all possible pairs of these nucleotides may be represented as the 4×4 matrix array in the partitioned form

$$\mathfrak{P} = \begin{pmatrix} AA & AG & AC & AT \\ GA & GG & GC & GT \\ CA & CG & CC & CT \\ TA & TG & TC & TT \end{pmatrix} = \begin{pmatrix} \mathfrak{P}_{11} & \mathfrak{P}_{12} \\ \mathfrak{P}_{21} & \mathfrak{P}_{22} \end{pmatrix}, \tag{6.8.2}$$

where, for example, the 2×2 matrix of symbols $\mathfrak{P}_{11}$ is defined by

$$\mathfrak{P}_{11} = \begin{pmatrix} AA & AG \\ GA & GG \end{pmatrix} \tag{6.8.3}$$

and the other matrices $\mathfrak{P}_{12}, \mathfrak{P}_{21}$ and $\mathfrak{P}_{22}$ are defined similarly. From the point of view of transitions and transversions the matrices $\mathfrak{P}_{11}$ and $\mathfrak{P}_{22}$ consist of transitions; while the matrices $\mathfrak{P}_{12}$ and $\mathfrak{P}_{21}$ consist of transversions. Given the array $\mathfrak{P}$ of symbols, the state space of the Markov process will be chosen as

$$\mathfrak{S} = unrav\,(\mathfrak{P})\,, \tag{6.8.4}$$

where the symbol $unrav(\mathfrak{P})$ means that the array $\mathfrak{P}$ is unraveled so that first row of $\mathfrak{P}$ becomes the first four elements of $\mathfrak{S}$, the next four elements of $\mathfrak{S}$ is the second row of $\mathfrak{P}$ and so on, which gives rise to representation of the set $\mathfrak{S}$ as an ordered array of 16 of dinucleotides. From now on the state space $\mathfrak{S}$ will be represented as the 1×16 array

$$\mathfrak{S} = \ (i \mid i = 1, 2, \ldots, 16) \tag{6.8.5}$$

of symbols.

Given this representation of $\mathfrak{S}$, the 16×16 rate matrix $\boldsymbol{Q}$ for dinucleotides transitions may be represented in the element by element form

$$\boldsymbol{Q} = \left(q_{ij} \mid i \in \mathfrak{S}; \mathrm{j} \in \mathfrak{S}\right), \tag{6.8.6}$$

where $q_{ij} \geq 0$ for all $i \neq j$ and

$$q_i = q_{ii} = -\sum_{j \neq i} q_{ij} \tag{6.8.7}$$

for every i. Therefore, just as for the rate matrices for Markov processes defined in previous sections, the condition

$$\sum_j q_{ij} = 0 \tag{6.8.8}$$

holds for every $i \in \mathfrak{S}$.

To further identify the rates in the matrix $\boldsymbol{Q}$, observe that the rate q_{12} is that for the transition $1 \to 2 = AA \to AG$. Observe that in this case, the nucleotide G was substituted for the nucleotide A in the second position of the dinucleotide AA. There may also be rates for dinucleotide substitutions involving both sites in a dinucleotide. For example, the rate q_{16} for dinucleotide substitution $1 \to 6 = AA \to GG$ is that for the case the substitution $A \to G$ occurs at each of the two sites. In general, for the case of 16×16 rate matrices, one may choose $16 \times 15 = 240$ free parameters, which would be a formidable task. It, therefore, seems prudent that when attempting to formulate a 16×16 matrix of rates that some principle be used to obtain a parsimonious definition of rates in the matrix $\boldsymbol{Q}$.

A simple and useful approach to formulating a 16×16 rate matrix is to generalize the method introduced by Felsenstein (1981). Let

$$\boldsymbol{\pi} = \left(\pi_i \mid i = 1, 2, \ldots, 16\right) \tag{6.8.9}$$

denote a 1×16 vector of probabilities of a stationary distribution. Then, use the following definitions to construct the elements of the 16×16 rate matrix $\boldsymbol{Q} = (q_{ij})$. For every $i \in \mathfrak{S}$, let $q_{ij} = v\pi_j$ if $j \neq i$, and if $j = i$, let $q_i = q_{ii} = -v\left(1 - \pi_i\right)$, where $v > 0$ is a parameter whose value is chosen to render the expected sojourn times in states plausible in terms of estimates of evolutionary times as discussed in foregoing sections. Given this method of construction, the rate matrix of Markov jump process in continuous time, it is easy to see that

$$\boldsymbol{\pi} = \boldsymbol{\pi}\boldsymbol{Q} = \mathbf{0}, \tag{6.8.10}$$

where $\mathbf{0}$ is a 1×16 vector of zeros. Therefore, $\boldsymbol{\pi}$ is the stationary distribution of the dinucleotide substitution process formulated as a Markov jump process in continuous time.

One of the problems that arise with this method of constructing the rate matrix $\boldsymbol{Q}$ is that of specifying the 16 elements of the stationary distribution $\boldsymbol{\pi}$. A useful approach to finding solutions to this problem is to suppose the dinucleotide substitution process governing the evolution of the n sites under consideration has been operative for tens of thousands of years so that the sample of DNA is in a statistical equilibrium governed by some stationary distribution $\boldsymbol{\pi}$. If it is assumed that this is the case, then one may proceed to estimate the stationary distribution. For example, let n_i denote the number of dinucleotides of type $i \in \mathfrak{S}$ in a sample of n dinucleotides, where

$$\sum_{i-1}^{16} n_i = n. \tag{6.8.11}$$

Then,

$$\widehat{\pi}_i = \frac{n_i}{n} \tag{6.8.12}$$

is an estimate of π_i for all $i \in \mathfrak{S}$. To complete the estimation process, let $\widehat{\upsilon}$ be a plausible estimate of the parameter υ that results in accepting expected sojourn times in state, and let $\widehat{\boldsymbol{Q}}$ be an estimate of the rate matix based on the procedure just described.

Then, $\widehat{\boldsymbol{P}}(t) = \exp\left(\widehat{\boldsymbol{Q}} \times t\right)$ would be an estimate of the transition matrix $\boldsymbol{P}(t)$. Let

$$\widehat{\boldsymbol{A}} = \mathbf{1}\widehat{\boldsymbol{\pi}}, \tag{6.8.13}$$

where $\mathbf{1}$ is a 16×1 vector of ones, and let

$$\widehat{\boldsymbol{R}}(t) = \widehat{\boldsymbol{P}}(t) - \widehat{\boldsymbol{A}}. \tag{6.8.14}$$

In principle, for ϵ small, one could program a computer to find a t_s such that

$$\| \widehat{\boldsymbol{R}}(t_s) \| = \epsilon, \tag{6.8.15}$$

which would yield an estimate of t_s, the time taken for the process to converge to a stationary distribution.

There is a special case that deserves some attention. Suppose the stationary distribution is the uniform distribution $\boldsymbol{\pi}_u$ so that

$$\boldsymbol{\pi}_u = \left(\pi_i = \frac{1}{16} \mid i = 1, 2, \ldots, 16\right). \tag{6.8.16}$$

Then, at equilibrium, not only would the dinucleotides be distributed independently but the nucleotides at single sites would also be distributed independently, and, in principle, statistical tests could be constructed to test for independence. If, however, the stationary distribution $\boldsymbol{\pi}$ is not uniform, then the dinucleotides in the sample of n sites of a DNA molecule would be independently distributed but the single sites would not be distributed independently when the population is in equilibrium. It is clear that the procedure just discussed could be extended to the case of tri-nucleotides or in a still more general cases to k-nucleotides for $k \geq 2$, but such extensions will not be considered here except to note that for the case of trinucleotides it would be necessary to construct 64×64 rate matrices.

6.9 A Retrospective View of a Markov Process with Stationary Transition Probabilities

In the foregoing sections, the evolution of a Markov process $X(t)$ in continuous time $t \geq 0$ with a finite state space $\mathfrak{S}$, rate matrix $\boldsymbol{Q}$ and transition matrix

$$\mathbf{P}(t) = (P_{ij}(t)) = \exp(\boldsymbol{Q}t) \tag{6.9.1}$$

defined for $t \geq 0$ was viewed in a forward or prospective direction. That is, given the state $X(s) = i \in \mathfrak{S}$ of the process at time $s. \geq 0$, attention was focused on the conditional probability that the process was in some state $X(s+t) = j \in \mathfrak{S}$ at a future time $s+t$. Under the assumption that the process has stationary transition probabilities, this conditional probability was given by

$$P[X(s+t) = j \mid X(s) = i] = P_{ij}(t). \tag{6.9.2}$$

In this section, reverse or retrospective conditional probabilities of the form

$$P[X(s) = i \mid X(s+t) = j] = P_{Rji}(t) \tag{6.9.3}$$

will be under consideration. Observe that it has been tacitly assumed that the transition probabilities for the reverse process are stationary in the sense that the probability on the right depends on the length of the backwards time interval $s+t-s = t$. To simplify the derivation of these retrospective conditional probabilities, it will be assumed all states in $\mathfrak{S}$ communicate so that the forward Markov process has a stationary distribution such that

$$\lim_{t \uparrow \infty} P_{ij}(t) = \pi_j \tag{6.9.4}$$

for every $i \in \mathfrak{S}$.

As a first step toward expressing the retrospective conditional probability $P_{Rji}(t)$ in terms of the forward conditional probability $P_{ij}(t)$, it will be helpful to observe that unconditional probability

$$P[X(s)=i, X(s+t)=j] \tag{6.9.5}$$

may be expressed in two ways; namely

$$\begin{aligned} & P[X(s)=i, X(s+t)=j] \\ &= P[X(s+t)=j]\,P[X(s)=i \mid X(s+t)=j] \\ &= P[X(s)=i]\,P[X(s+t)=j \mid X(s)=i]. \end{aligned} \tag{6.9.6}$$

Therefore,

$$P[X(s+t)=j]\,P_{Rji}(t) = P[X(s)=i]\,P_{ij}(t). \tag{6.9.7}$$

By letting $s \uparrow \infty$ in this equation, it follows that

$$\pi_j P_{Rji}(t) = \pi_i P_{ij}(t). \tag{6.9.8}$$

From this result, it can be seen that when the forward process is in statistical equilibrium, the transition probabilities for the retrospective process are given by

$$P_{Rji}(t) = \pi_i P_{ij}(t)\,\pi_j^{-1} \tag{6.9.9}$$

for all pairs of states i and j. It is also of interest to note that the retrospective conditional probabilities satisfy the condition

$$\sum_i P_{Rji}(t) = \left(\sum_i \pi_i P_{ij}(t)\right)\pi_j^{-1} = 1 \tag{6.9.10}$$

for every $j \in \mathfrak{S}$.

Let

$$\boldsymbol{P}_R(t) = (P_{Rji}(t)) \tag{6.9.11}$$

denote a matrix of retrospective probabilities for the reverse process. To express this matrix in terms of the forward matrix $\boldsymbol{P}(t)$, let

$$\boldsymbol{\Pi} = \boldsymbol{diag}(\pi_i \mid i \in \mathfrak{S}) \tag{6.9.12}$$

denote a diagonal matrix such that the stationary probabilities are on the principal diagonal. Then, it is easy to see that

$$\boldsymbol{P}_R(t) = \boldsymbol{\Pi}\boldsymbol{P}(t)\boldsymbol{\Pi}^{-1} = \boldsymbol{\Pi}\exp(\boldsymbol{Q}t)\boldsymbol{\Pi}^{-1}. \tag{6.9.13}$$

From this expression, it is also easy to see that the risk matrix for the retrospective process is given by

$$\boldsymbol{Q}_R = (q_{Rji}) = \boldsymbol{\Pi}\boldsymbol{Q}\boldsymbol{\Pi}^{-1}. \tag{6.9.14}$$

A concept that is related to the reverse process is the idea of a reversible Markov process in continuous time. A process will be said to be reversible if

$$P_{Rji}\left(t\right) = P_{ji}\left(t\right) \tag{6.9.15}$$

and

$$\pi_j P_{ji}(t) = \pi_i P_{ij}\left(t\right) \tag{6.9.16}$$

for all pairs of states i, j and $t > 0$. From this definition, it is easy to see that if the rate matrix $\boldsymbol{Q}$ is symmetric and there are $r \geq 2$ states in $\mathfrak{S}$, then this equation would hold for all pairs of states because the stationary distribution of the process is the uniform distribution, where

$$\pi_i = \frac{1}{r} \tag{6.9.17}$$

for all $i \in \mathfrak{S}$ and $P_{ji}(t) = P_{ij}(t)$ for all pairs i, j and $t \geq 0$. Therefore, if a process has a symmetric rate matrix $\boldsymbol{Q}$, then it is reversible. It is recommended that if a reader is interested in more detailed discussion of reversible Markov chains in discrete time, the book, Ewens and Grant (2005), may be consulted.

It is also of interest to provide an outline of a procedure that may be used to calculate a numerical version of a retrospective distribution of a process, given that the forward process is in statistical equilibrium and some state j is observed at time $s > 0$. The first step in such a procedure is to obtain a numerical version of the rate matrix $\boldsymbol{Q}$. Let $\widehat{\boldsymbol{Q}}$ denote this numerical version. The elements of this matrix may be obtained in at least two ways. One could, for example, assign plausible values to the elements in this matrix in accordance with some theory under consideration, or if data were available, one could estimate the elements of the rate matrix $\boldsymbol{Q}$. Then, given a numerical version of the rate matrix $\widehat{\boldsymbol{Q}}$, the matrix of forward transition probabilities

$$\widehat{\boldsymbol{P}}\left(t\right) = \exp\left(\widehat{\boldsymbol{Q}}t\right) \tag{6.9.18}$$

could be calculated as a function of $t \geq 0$. Such software packages such as MATLAB or MAPLE could be used to carry out these calculations. In particular, if one used MAPLE or some other symbolic computation software, it may be possible to obtain symbolic versions of the matrix $\widehat{\boldsymbol{P}}\left(t\right)$.

The next step in the procedure could be that of calculating the stationary distribution of the process. If the state space $\mathfrak{S}$ contains $r \geq 2$ states, let

$$\widehat{\boldsymbol{\pi}} = \left(\widehat{\pi}_1, \widehat{\pi}_2, \ldots, \widehat{\pi}_r\right) \tag{6.9.19}$$

denote a $1 \times r$ vector of stationary probabilities calculated from the numerical version of the rate matrix $\widehat{\boldsymbol{Q}}$. Briefly, the vector $\widehat{\boldsymbol{\pi}}$ would be a normalized left eigenvector corresponding to the eigenvalue $\widehat{\rho} = 0$ of the matrix $\widehat{\boldsymbol{Q}}$. Any software package with programs designed to find eigenvectors corresponding to given eigenvalues could be used to do this calculation. As will be shown below, the calculations just described are sufficient to determine a numerical version of the retrospective distribution.

For example, suppose one wished to find the retrospective distribution of the process for t units of time in the past, given that the forward process was in statistical equilibrium and the observed present state of the process was $j \in \mathfrak{S}$. Then, from the foregoing discussion, it follows that the numerical version of the retrospective distribution is

$$\widehat{P}_{Rji}(t) = \left(\widehat{\pi}_i \widehat{P}_{ij}(t)\right) \widehat{\pi}_j^{-1} \tag{6.9.20}$$

for all $i \in \mathfrak{S}$, where $\widehat{P}_{ij}(t)$ is element i, j in the matrix $\widehat{\boldsymbol{P}}(t) = \exp\left(\widehat{\boldsymbol{Q}}t\right)$. Given this formula, a retrospective conditional probability could be evaluated for selected values of $t > 0$ at each state in $\mathfrak{S}$.

Numerous other retrospective conditional probabilities may also be expressed in terms of a forward transition probabilities and elements of the stationary distribution. For example, suppose, given that time s the process was in state i and at time $s + t$ it was in state j but no where in the open time interval $(s, s + t)$ was the process in state j. Let $F_{Rji}(t)$ denote the retrospective conditional probability of this event. With respect to the forward process, let $F_{ij}(t)$ denote the conditional probability that, given $X(s) = i$, the process enters state j for the first time during the time interval $(s, s + t]$. Then, if the forward process is in statistical equilibrium, it follows that

$$F_{Rji}(t) = (\pi_i F_{ij}(t)) \pi_j^{-1}. \tag{6.9.21}$$

Given this formula, a problem that presents itself is that of developing a procedure to evaluate the righthand side numerically. As will be shown, it will be possible to express the function $F_{ij}(t)$ as an element of an exponential matrix after the rate matrix $\boldsymbol{Q}$ has been properly rearranged. To this end, let $\mathfrak{S}_1 = (j)$ denote a set that contains only the state j, and let $\mathfrak{S}_2 = \mathfrak{S}_1^c$, the complement of $\mathfrak{S}_1$. Then, construct a $r \times r$ partitioned rate matrix from the rate matrix $\boldsymbol{Q}$ of the form

$$\boldsymbol{Q}^* = \begin{pmatrix} 0 & \mathbf{0} \\ \boldsymbol{Q}_{21}^* & \boldsymbol{Q}_{22}^* \end{pmatrix}, \tag{6.9.22}$$

where $\mathbf{0}$ is a $1 \times (r-1)$ vector of zeros, $\boldsymbol{Q}_{21}^{*}$ is a $(r-1) \times 1$ vector of rates and $\boldsymbol{Q}_{22}^{*}$ is a $(r-1) \times (r-1)$ matrix of rates. In this partitioned matrix, the elements of $\boldsymbol{Q}_{21}^{*}$ are rates governing transitions from states in $\mathfrak{S}_2$ to the state j and the elements of $\boldsymbol{Q}_{22}^{*}$ are rates governing transitions among the states in the set $\mathfrak{S}_2$. Thus, elements of the matrices $\boldsymbol{Q}_{21}^{*}$ and $\boldsymbol{Q}_{22}^{*}$ are merely rearrangements of the elements of the rate matrix $\boldsymbol{Q}$, and the first row of $\boldsymbol{Q}^{*}$ indicates the state j has been transformed to an absorbing state in a modified Markov process with one absorbing state.

Let

$$\boldsymbol{P}^{*}(t) = \left(P_{ij}^{*}(t)\right) = \exp\left(\boldsymbol{Q}^{*}t\right) \tag{6.9.23}$$

denote a matrix of transition probabilities computed from the modified rate matrix $\boldsymbol{Q}^{*}$. Then, for $i \neq j$, it follows that

$$F_{ij}(t) = P_{ij}^{*}(t)\,. \tag{6.9.24}$$

This function could, therefore, be computed given any software package that contains a program to compute the exponential matrix. Two such software packages are MATLAB and MAPLE.

It can be shown that $F_{ij}(t)$ is a non-decreasing function of $t \geq 0$ and

$$\lim_{t\uparrow\infty} F_{ij}(t) = 1. \tag{6.9.25}$$

Therefore, $F_{ij}(t)$ may interpreted as the distribution function of a random variable T_{ij}, representing the waiting time for the forward process to enter state j for the first time at some point in the interval of length $t > 0$, given that the process was in state $i \neq j$ and at some initial time $t = 0$. Let $f_{ij}(t) = F_{ij}^{'}(t)$ denote the density of $F_{ij}(t)$. Then, a quantity of interest would be the expectation

$$E\left[T_{ij}\right] = \int_{0}^{\infty} t f_{ij}(t)\,dt \tag{6.9.26}$$

of the time for the forward process to enter j for the first time, given that at some initial time $t = 0$ the process was in state $i \neq j$. Methods for finding numerical values of these expectations will not be pursued here, but it is suggested that an interested reader consult chapter 7 of Mode and Sleeman (2000), where Markov jump processes in continuous time with one or more absorbing states are viewed from a semi-Markov perspective. This perspective leads to linear renewal integral equations and a straight forward use of LaPlace transforms reduces the problem of finding expectations of the form $E\left[T_{ij}\right]$, as well as other expectations, as solutions of linear algebraic equations depending on the parameters of the model.

Bibliography

[1] Bellman, R. (1953) **Stability Theory of Differential Equations**. McGraw-Hill, New York, Toronto, London.

[2] Bellman, R. (1960) **Introduction to Matrix Analysis**. McGraw-Hill, New York, Toronto, London.

[3] Blaisdell, B. (1985) A method for estimating from two aligned present day DNA sequences their ancestral composition and subsequent rates of substitution. J. Mol. Evolution **22**:69–81.

[4] Capasso, V. (1991) **Mathematical Structures of Epidemic Systems, Lecture Notes in Biomathematics Vol. 97**. Springer-Verlag. Berlin, New York, London, Paris.

[5] Ewens, W. J. and Grant, G. R. (2005). **Statistical Methods in Bioinformatics - An Introduction, Second Edition**, Springer.

[6] Feller, W. (1968) **An Introduction to Probability Theory and Its Applications. Volume I, Third Edition**. John Wiley and Sons, New York, London, Sydney.

[7] Feller, W. (1966) **An Introduction to Probability Theory and Its Applications. Volume II**, John Wiley and Sons, New York, London, Sydney.

[8] Felsenstein, J. (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. **17**:368–376.

[9] Felsenstein, J. and Churchill, G. A. (1996) A hidden Markov model approach to variation among sites in rate of evolution. Mol. Biol. Evol. **13**:93–104.

[10] Hasegawa, M., Kishino, H. and Yano, T. (1985) Dating human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22**:160–174.

[11] Jukes, T. H. and Cantor, C. R. (1969) Evolution of Protein Molecules. pp. 21-132 in **Mammalian Protein Metabolism**, H. N. Monro, Editor, Academic, Press New York.

[12] Karlin, S. and Taylor, H. M. (1975) **A First Course in Stochastic Processes**. Academic Press, Boston, New York, London. (second edition)

[13] Karlin, S. and Taylor, H. M. (1981) **A Second Course in Stochastic Processes**. Academic Press, Boston, New York, London.

[14] Kimura, M. (1980) A Simple Method for Estimating Evolutionary Rate in a Finite Population Due to Mutational Production of Neutral and Nearly Neutral Base Substitutions Through Comparative Studies of Nucleotide Sequences. J. Molec. Biol **16**:111–120.

[15] Li, W. H. (1997) **Molecular Evolution**. Sinauer Associates Inc. Sunderland, Mass 01375.

[16] Mode, C. J. and Sleeman, C. K. (2000) **Stochastic Processes in Epidemiology, HIV/AIDS, Other Infectious Diseases and Computers**. World Scientific, Singapore, New Jersey, London, Hong Kong.

[17] Nei, M. and Kumar, S. (2000) **Molecular Evolution and Phylogenetics**. Oxford University Press.

[18] Popper, K. R. (1939) **The Logic of Scientific Discovery**. Basic Books, Inc., New York.

[19] Strachan, T. and Read, A. P. (2004) **Human Molecular Genetics, Third Edition**. Garland Science, Taylor and Francis Group, London and New York.

[20] Tavaré, S. (1986) **Lectures on Mathematics in the Life Sciences**. **17**:57–86.

[21] Yang, Z. (2006) **Computational Molecular Evolution**. Oxford University Press.

Chapter 7

Mixtures of Markov Processes as Models of Nucleotide Substitutions at Many Sites

7.1 Introduction

As one reads the existing literature on models of nucleotide substitution, such as the Markov models described in chapter 6, it is sometimes stated that it is thought that rates of substitution may vary among sites and that rates at adjacent sites may be correlated. As was pointed out in that chapter, some investigators have also considered Markov models that accommodate position dependence, i.e., the notion that substitution rates at a particular site may depend on the nucleotides present at adjacent sites of a DNA molecule. As has been stated in previous chapters in which attention was focused on Wright-Fisher models, the sequencing of the human genome and that of other species has uncovered the existence of single nucleotide polymorphisms throughout the genomes investigated, which suggests that the evolutionary process of nucleotide substitution has been occurring at literally millions of sites in the genomes of several species.

Therefore, when contemplating the modeling of the evolutionary process of nucleotide substitution with respect to many sites of large DNA molecules using the techniques described in chapter 6, one comes face to face with the daunting problem of considering thousands of parameters that go into the specification of large rate matrices governing Markov processes. It thus becomes clear that if one wishes to entertain models of nucleotide substitution with respect to a large number of sites, it will be necessary to shift the working paradigm from the formal mathematics considered in chapter 6 to computer intensive methods based on some well defined mathematical structure.

Briefly, the mathematical structure considered and implemented in this chapter consists of two components. One component is a stationary process, depending on relatively few parameters, that is used to generate rates of nucleotide substitutions that lie in subintervals of $(0, 1)$. These subintervals in turn are chosen in accordance with some prior ideas about plausible domains for rates per unit time, governing the evolutionary processes of nucleotide substitutions at many sites. This stationary process accommodates the notion that rates vary and may be correlated among sites. Such correlations may also contribute to the perception that rates of nucleotide substitutions are position dependent. A second component of the structure is a four-state Markov substitution process at each site under consideration.

If one were to work within the paradigm described in chapter 6, it would be necessary to generate 12 free parameters for the rate matrix at every site. In computer implementations of this mathematical structure, however, some parsimony is attained by using the idea that if a nucleotide occupies a site at a given time, then only three nucleotide substitutions may occur; namely to one of three nucleotides that are different from the original one. Thus, if n sites are under consideration, then it suffices to generate only $3 \times n$ rates rather than the $12 \times n$ rates that would be necessary if the one worked within the paradigm described in chapter 6.

Just as in chapter 6, different rates for transitions and transversions are accommodated in the model. In the remaining sections of this chapter, the technical details entailed in the construction of this two component model will be discussed in depth. In particular, attention will be focused on issues centering around the choice of a stationary Gaussian process such that it will be feasible to compute long arrays of numbers which may be mapped into subintervals of the interval $(0, 1)$ representing rates of nucleotide substitutions at many sites of a DNA molecule. A second set of issues will deal with problems that are encountered when one wishes to map uniform random variables on the interval $(0, 1)$ to random variables with some specified distribution.

7.2 Mixtures of Markov Models and Variable Substitution Rates Across Sites

In applications of the one parameter Jukes-Cantor model or of the two parameter Kimura model some authors, see, for example, Nei and Kumar (2000), have used the idea that parameters in a rate matrix are realizations of a random variable $\boldsymbol{\Theta}$ with some distribution on the interval $(0, \infty)$. To illustrate this concept, suppose the rate matrix $\boldsymbol{Q}(\theta)$ is a function of some parameter $\theta \in \mathbb{R}^+ = \ (0, \infty)$ and this random variable has the density function $g(\theta)$. Furthermore, suppose that at time $t = 0,$ the initial nucleotide is $i \in \mathfrak{S}$ and that the nucleotide substitution process evolves as a Markov process in continuous time, given a realization of the random variable $\mathbf{\Theta} = \theta$. Let $P_{ij}\left(\theta.t\right)$ denote the conditional probability that the process is in state $j \in \mathfrak{S}$ at time $t > 0$, given $\mathbf{\Theta} = \theta$ and the condition that process was in state i at time $t = 0$. In symbols, let the random function $X\left(t\right)$ denote the state of the Markov process at time t. Then, formally,

$$P_{ij}\left(\theta.t\right) = P\left[Xt = j \mid \mathbf{\Theta} = \theta, X\left(0\right) = i\right]. \tag{7.2.1}$$

By definition, the unconditional probability that the process is in state j at time $t > 0$, given the initial state i, is

$$P_{ij}\left(t\right) = P\left[X\left(t\right) = j \mid X\left(0\right) = i\right] = \int_0^\infty g\left(\theta\right) P_{ij}\left(\theta.t\right) d\theta. \tag{7.2.2}$$

For the case of one site, this equation defines what will be referred as a mixture of Markov processes, where $g\left(\theta\right)$ is called the mixing distribution. It is of interest to note, that conditional on $\mathbf{\Theta} = \theta$, the process

$$\left(X\left(\theta, t\right) \mid t \in \mathbb{T}^+, \theta \in \mathbb{R}^+\right) \tag{7.2.3}$$

has the Markov property by construction for every $\theta \in \mathbb{R}^+$, but the unconditional process

$$\left(X\left(t\right) \mid t \in \mathbb{T}^+\right) \tag{7.2.4}$$

will not in general satisfy the Markov property. The parametric form chosen for the mixing density function $g\left(\theta\right)$ by several authors, see Nei and Kumar (2000), is the well known gamma distribution with the formula

$$g\left(\theta\right) = \frac{\beta^\alpha}{\Gamma\left(\alpha\right)} \theta^{\alpha-1} \exp\left(-\beta\theta\right) \text{ for } \theta \in \left(0, \infty\right), \tag{7.2.5}$$

where $\alpha > 0$ and $\beta > 0$ are parameters and $\Gamma\left(\cdot\right)$ is the gamma function. As is widely known, the expectation and variance of the random variable $\mathbf{\Theta}$ are

$$E\left[\mathbf{\Theta}\right] = \int_0^\infty \theta g\left(\theta\right) d\theta = \frac{\alpha}{\beta} \tag{7.2.6}$$

and

$$var\left[\mathbf{\Theta}\right] = E\left[\left(\mathbf{\Theta} - \ E\left[\mathbf{\Theta}\right]\right)^2\right] = \frac{\alpha}{\beta^2}. \tag{7.2.7}$$

An interesting property of the idea of using mixtures of Markov processes as models for the evolutionary process of nucleotide substitution is that it provides a framework of generalizing from one site to many sites with different rates of substitution at each site. As an illustration, suppose it is desired to construct a model of nucleotide substitution process for $n \geq 1$ base sites, where n is a large number. For every site $\nu = 1, 2, \ldots, n$ and $t > 0$, let $P_{i_\nu j_\nu}(\theta_\nu, t)$ denote the conditional probability that the process at the *v-th* site is in state j_ν at time t, given that the random variable $\boldsymbol{\Theta}_\nu$ $= \theta_\nu$ and the initial state at $t = 0$ was i_ν.

To formally define a stochastic process for $n \geq 1$ sites, it will be helpful to define some product sets Let $\mathbb{R}_n^+$ denote the n-fold Cartesian product of the set $\mathbb{R}^+$ with itself. In symbols

$$\mathbb{R}_n^+ = ((\theta_1, \theta_2, \ldots, \theta_n) \mid \theta_\nu \in (0, \infty) \text{ for all } \nu = 1, 2, \ldots, n). \qquad (7.2.8)$$

Similarly, let $\mathfrak{S}_n$ denote the n-fold Cartesian product of the set $\mathfrak{S}$ with itself. Formally,

$$\mathfrak{S}_n = ((i_1, i_2, \ldots, i_n) \mid i_\nu \in \mathfrak{S} \text{ for all } \nu = 1, 2, \ldots, n). \qquad (7.2.9)$$

The sets $\mathbb{R}_n^+$ and $\mathfrak{S}_n$ may also be viewed as sets of n-dimensional spaces of vectors. It will also be useful in what follows to let the symbol $\mathbb{S}$ denote the set of n sites under consideration.

Next, let

$$\boldsymbol{\Theta} = (\boldsymbol{\Theta}_\nu \mid \nu = 1, 2, \ldots, n) \qquad (7.2.10)$$

denote a n-dimensional vector random variable taking values in the product set $\mathbb{R}_n^+$, and for

$$\boldsymbol{\Theta} = \boldsymbol{\theta} = (\theta_\nu \mid \nu = 1, 2, \ldots, n) \qquad (7.2.11)$$

fixed, let

$$\boldsymbol{X}(\boldsymbol{\theta}, t) = (X_\nu(\theta_\nu, t) \mid \nu = 1, 2, \ldots, n) \qquad (7.2.12)$$

be a vector valued random function defined for $t \in \mathbb{T}^+$ and taking values in the product set $\mathfrak{S}_n$.

Given these definitions, the assumption that characterizes the evolution of the process among the n sites is expressed in the following statements and equation. For every pair of vectors $\boldsymbol{i} = (i_1, i_2, \ldots, i_n)$ and $\boldsymbol{j} = (j_1, j_2, \ldots, , j_n)$ in $\mathfrak{S}_n$, it will be assumed that

$$P[\boldsymbol{X}(\boldsymbol{\theta}, t) = \boldsymbol{j} \mid \boldsymbol{\Theta} = \boldsymbol{\theta}, \boldsymbol{X}(\boldsymbol{\theta}, 0) = \boldsymbol{i}] = \prod_{\nu=1}^{n} P_{i_\nu j_\nu}(\theta_\nu, t). \qquad (7.2.13)$$

for all $t \in \mathbb{T}^+$. In other words, given $\boldsymbol{\Theta} = \boldsymbol{\theta} \in \mathbb{R}_n^+$, it is assumed that the evolution of the Markov processes among the n sites are conditionally independent. It will also be assumed that for any subset S_k of $\mathbb{S}$ for $k \geq 2$ this assumption of conditional independence is also in force.

To complete the definition of a class of nucleotide substitution processes for n sites, it will be necessary to define the distribution of the $\boldsymbol{\Theta}$-process for every subset S_k of $\mathbb{S}$ with k elements such that $1 \leq k \leq n$. For every subset S_k, let $\mathbb{R}^+_{S_k}$ denote the subspace of $\mathbb{R}^+_n$ indexed by the elements of the subset S_k and let $g\left(\boldsymbol{\theta}_{\mathbb{S}_k}\right)$ denote the density function of the vector random variable $\boldsymbol{\Theta}_{S_k}$ with values in $\mathbb{R}^+_{S_k}$. Let

$$\mathcal{F} = \left(g\left(\boldsymbol{\theta}_S\right) \mid S \subset \mathbb{S}\right) \tag{7.2.14}$$

denote the family of all such densities, where S ranges over all non-empty subsets of $\mathbb{S}$. The family $\mathcal{F}$ will be said to be consistent if for every subset S_k of $\mathbb{S}$ and every subset U_l of S_k such that $l < k$, the equation

$$g\left(\boldsymbol{\theta}_{U_l}\right) = \int_{\mathbb{R}^+_{k-l}} g\left(\boldsymbol{\theta}_{S_k}\right) d\boldsymbol{\theta}_{k-l} \tag{7.2.15}$$

is satisfied. In other words, this equation states that the density corresponding to indices in the subset U_l of S_k is the same for all k such that $k > l$.

Among the advantages to the consideration of mixtures of Markov models of nucleotide substitution on n sites is that at statistical equilibrium the nucleotides are not distributed independently among sites as was the case for unconditioned Markov processes. To see that this statement is indeed true for the mixture process under consideration, for every pair of vectors $\boldsymbol{i}$ and $\boldsymbol{j}$ in $\mathfrak{S}_n$ let

$$P\left[\boldsymbol{X}(t) = \boldsymbol{j} \mid \boldsymbol{X}(0) = \boldsymbol{i}\right] \tag{7.2.16}$$

denote the unconditional probability that the mixture process is in state $\boldsymbol{j} \in \mathfrak{S}_n$ at time $t > 0$, given it was in state $\boldsymbol{i}$ at time $t = 0$. Then, because of the assumption of conditional independence, given $\boldsymbol{\Theta} = \boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$, it follows that

$$P\left[\boldsymbol{X}(t) = \boldsymbol{j} \mid \boldsymbol{X}(0) = \boldsymbol{i}\right] = \int_{\mathbb{R}_n} g(\boldsymbol{\theta}) \prod_{\nu=1}^{n} P_{i_\nu j_\nu}\left(\theta_\nu, t\right) d\boldsymbol{\theta}. \tag{7.2.17}$$

However, for every $\nu = 1, 2, \ldots, n$ and $\theta_\nu \in \mathbb{R}^+$

$$\lim_{t \uparrow \infty} P_{i_\nu j_\nu}\left(\theta_\nu, t\right) = \pi_{j_\nu}\left(\theta_\nu\right), \tag{7.2.18}$$

where $\pi_{j_\nu}\left(\theta_\nu\right)$ is the stationary probability for state $j_\nu \in \mathfrak{S}$, given θ_ν for site ν. Let

$$\boldsymbol{\pi} = \left(\pi_{j_1}, \pi_{j_2}, \ldots, \pi_{j_n}\right) \tag{7.2.19}$$

denote the unconditional stationary distribution of the mixture process. By definition, this distribution is

$$\boldsymbol{\pi} = \lim_{t\uparrow\infty} P\left[\boldsymbol{X}(t) = \boldsymbol{j} \mid \boldsymbol{X}(0) = \boldsymbol{i}\right] = \int_{\mathbb{R}_n} g(\boldsymbol{\theta}) \prod_{\nu=1}^{n} \pi_{j_\nu}(\theta_\nu)\, d\boldsymbol{\theta}. \tag{7.2.20}$$

This integral shows that, in general,

$$\boldsymbol{\pi} \neq \prod_{\nu=1}^{n} \pi_{j_\nu}. \tag{7.2.21}$$

Therefore, at statistical equilibrium for the mixture process, the nucleotides at the n sites are not distributed independently as was the case for unconditioned Markov processes. Similar sets of equations could be written down for the marginal distributions corresponding to every subset S_k of $\mathbb{S}$ such that $2 \leq k \leq n$ but the details will be omitted.

Observe that the condition of different evolutionary rates at each site follows from properties of the mixture process, because the rate matrix $\boldsymbol{Q}_\nu(\theta_\nu)$ of this process at every site $\nu = 1, 2, \ldots, n$ depends on one or more realizations of the random variable $\boldsymbol{\Theta}_\nu$. It is in this sense that the rates of evolution among the n sites may differ. In applying mixed Markov processes to the evolution of nucleotide substitutions at n sites for two related species, the time $t = 0$ may be interpreted as the time the two species separated. Under this interpretation the initial nucleotide sequence

$$\boldsymbol{i} \;=\; (i_\nu \mid \nu = 1, 2, \ldots, n) \tag{7.2.22}$$

would be the sequence that was presumed to be present before the two species separated. In the next section, some examples of the $\boldsymbol{\Theta}$-process will be described and analyzed.

7.3 Gaussian Mixing Processes

A direct approach to constructing a mixing process based on a consistent family of distributions, described in the previous section, is to start with a Gaussian process and then map this process to intervals in $(0, 1)$ that seem plausible values for nucleotide substitution rates over a set of nucleotide sites governed by Markov processes with the common state space $\mathfrak{S}$. As in the foregoing section, let $\mathbb{S}$ denote the set of $n \geq 1$ nucleotide sites under consideration and let s_1 and s_2 be any two sites in $\mathbb{S}$. It will be supposed that these sites are numbered from 1 to n. As a first step in constructing a Gaussian process $\mathbf{Z}(s)$ on the set $s \in \mathbb{S}$ of sites, it will be necessary to define a mean or expectation function $E\left[\mathbf{Z}\left(s\right)\right]=\mu\left(s\right)$ for all $s\in\mathbb{S}$ and a covariance function

$$E\left[\left(\mathbf{Z}\left(s_1\right)-\mu\left(s_1\right)\left(\mathbf{Z}\left(s_2\right)-\mu\left(s_2\right)\right)\right]=\gamma\left(s_1,s_2\right) \tag{7.3.1}$$

for every pair of sites s_1 and s_2 in $\mathbb{S}$. If $s_1=s_2=s$, then $\gamma\left(s,s\right)$ is the variance of the random variable $\mathbf{Z}\left(s\right)$ for all $s\in\mathbb{S}$.

In order for the function $\gamma\left(s_1,s_2\right)$ to be a covariance function, it is necessary that it satisfy the condition of non-negative definiteness, which is defined as follows. Let

$$\mathbf{\Gamma}_k=\left(\gamma\left(s_i,s_j\right)\mid i,j=1,2,\ldots,k\right) \tag{7.3.2}$$

denote $k\times k$ covariance matrix corresponding to any subset $\mathbb{S}_k$ of $\mathbb{S}$ with k sites for $1\leq k\leq n$. The covariance function $\gamma\left(s_i,s_j\right)$ will have the non-negative definite property if for every $k\times 1$ column vector

$$\boldsymbol{x}_k=\left(x_i\mid i=1,2,\ldots,k\right) \tag{7.3.3}$$

of real numbers, the inequality

$$\boldsymbol{x}_k^T\mathbf{\Gamma}_k\boldsymbol{x}_k=\sum_{i=1}^{k}\sum_{j=1}^{k}\gamma\left(s_i,s_j\right)x_ix_j\geq 0 \tag{7.3.4}$$

is satisfied for every subset $\mathbb{S}_k$ of $\mathbb{S}$ such that $1\leq k\leq n$.

In books on mathematical probability theory that treat the subject of characteristic functions, it is shown that every characteristic function of a random variable has the property of non-negative definiteness, see for example the books of Loève (1955) and Lukacs (1960), where this property is listed under a result called Bochner's theorem. By definition the characteristic function of a random variable Z with density $f\left(z\right)$ is the expectation

$$\phi\left(u\right)=E[\exp(iuZ)]=\int_{-\infty}^{\infty}\exp\left(iuz\right)f\left(z\right)dz \tag{7.3.5}$$

defined for all $u\in\left(-\infty,\infty\right)$, where $i=\sqrt{-1}$, the imaginary element such that $i^2=-1$. In general, $\phi\left(u\right)$ is a complex valued function of a real variable u, but in some cases $\phi\left(u\right)$ is real valued.

A commonly used concept in books on mathematical statistics is a function called the moment generating function, which is defined by the expectation

$$m\left(t\right)=E[\exp(tZ)]=\int_{-\infty}^{\infty}\exp\left(tz\right)f\left(z\right)dz \tag{7.3.6}$$

for those $t\in\left(-\infty,\infty\right)$ for which the integral converges. For example, if a random variable X has a normal distribution with expectation μ and variance σ^2, then it can be shown that the moment generating function of X is

$$m(t) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp(tx) \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx = \exp\left(t\mu + \frac{\sigma^2}{2}t^2\right) \tag{7.3.7}$$

for all $t \in (-\infty, \infty)$. The book Bain and Englehardt (1987) on mathematical statistics may be consulted for details. If $m(t)$ is the moment generating function of a random variable Z, then its characteristic function is $\phi(u) = m(iu)$. Therefore, the characteristic function of a normal random variable X with expectation μ and variance σ^2 is

$$\phi(u) = \exp\left(iu\mu - \frac{\sigma^2}{2}u^2\right). \tag{7.3.8}$$

In particular, if $\mu = 0$, then $\phi(u)$ has the real form

$$\phi(u) = \exp\left(-\frac{\sigma^2}{2}u^2\right) \tag{7.3.9}$$

and is thus a candidate for a covariance function of a stochastic process. When viewing $\phi(u)$ as a covariance function the symbol σ^2 will be replaced by $\beta > 0$. By applying Bochner's theorem, one can conclude that the function

$$h(s_1, s_2) = \exp\left(-\frac{\beta}{2}(s_2 - s_1)^2\right), \tag{7.3.10}$$

which is defined for all pairs (s_1, s_2) of points in $\mathbb{S}$, has the property of non-negative definiteness. Furthermore, with a view towards constructing a Gaussian process with a minimal number of parameters, it will be supposed that for every $s \in \mathbb{S}$ the variance of the random variable $\boldsymbol{Z}(s)$ is $\sigma^2 > 0$. The function $\gamma(s_1, s_2)$ defined by

$$\gamma(s_1, s_2) = \sigma^2 \exp\left(-\frac{\beta}{2}(s_2 - s_1)^2\right) \tag{7.3.11}$$

will, therefore, be selected as the covariance function of the Gaussian process $\mathbf{Z}(s)$ on the set $s \in \mathbb{S}$ of nucleotide sites. Observe that this is an example of a stationary covariance function, because it depends only on the difference $s_2 - s_1$. Finally, to complete the construction of the Gaussian process $\mathbf{Z}(s)$ on the set $s \in \mathbb{S}$, it will be assumed that for every $s \in \mathbb{S}$, the expectation of the random function $\mathbf{Z}(s)$ is the constant

$$E[\mathbf{Z}(s)] = \mu. \tag{7.3.12}$$

This construction involves only three parameters μ, σ^2 and β and thus meets the criterion of a Gaussian process with a minimal number of parameters. It is also interesting to observe that this process is stationary.

The correlation function of the process is, by definition,

$$\rho\left(s_1, s_2\right)=\frac{\gamma\left(s_1, s_2\right)}{\sqrt{\gamma\left(s_1, s_1\right) \times \gamma\left(s_2, s_2\right)}}=\exp \left(-\frac{\beta}{2}\left(s_2-s_1\right)^2\right) \tag{7.3.13}$$

for all pairs (s_1, s_2) of sites in $\mathbb{S}$. For every value of $\beta > 0$, this function decreases as the difference $\mid s_2 - s_1 \mid$ increases, and for values of $\beta \geq 1$, this decrease may be rapid so that rates of nucleotide substitution at sites far apart will not be highly correlated. However, for some values of β such that $0 < \beta < 1$ correlations among rates of substitutions at all n sites may be quite high. For example, suppose $\beta = 1/n^2$, then

$$\frac{\beta}{2}\left(s_2-s_1\right)^2=\frac{\left(s_2-s_1\right)^2}{2 n^2} \leq \frac{1}{2} \tag{7.3.14}$$

for all pairs (s_1, s_2). Therefore,

$$-\frac{1}{2} \leq-\frac{\beta}{2}\left(s_2-s_1\right)^2 \leq 0, \tag{7.3.15}$$

which implies that

$$\exp \left(-\frac{1}{2}\right) \leq \exp \left(-\frac{\beta}{2}\left(s_2-s_1\right)^2\right)=\rho\left(s_1, s_2\right) \leq 1 \tag{7.3.16}$$

for all pairs (s_1, s_2). However, $\exp\left(-\frac{1}{2}\right) = 0.606530659712633$ so that

$$0.606530659712633 \leq \rho\left(s_1, s_2\right) \leq 1 \tag{7.3.17}$$

for all pairs (s_1, s_2), which shows that when $\beta = 1/n^2$ correlations among rates of nucleotides substitutions can be quite high among all $n \geq 2$ sites under consideration.

As mentioned in the foregoing section, the mixing process for the substitutions rates of the n Markov processes under consideration must be a consistent family of distributions. To see that the Gaussian process just constructed does satisfy the consistency condition, it will be helpful to describe the characteristic function of a multi-dimensional normal distribution. Let

$$\boldsymbol{Z}=(Z(s) \mid s=1,2, \ldots, n) \tag{7.3.18}$$

denote a $n \times 1$ column vector of the random variables of the mixing Gaussian process. As above, let

$$\mathbf{\Gamma}_n=\left(\gamma\left(s_i, s_j\right) \mid i=1,2, \ldots, n ; j=1,2, \ldots, n\right) \tag{7.3.19}$$

denote the $n \times n$ covariance matrix of the vector random variable $\boldsymbol{Z}$. Next let

$$\boldsymbol{\mu}_n = E\left[\boldsymbol{Z}\right] = \mu \mathbf{1}_n \tag{7.3.20}$$

denote the expectation vector of the random vector $\boldsymbol{Z}$, where $\mathbf{1}_n$ is a $n \times 1$ column vector such that each element is the number one. Finally, let $\boldsymbol{u}$ denote a $n \times 1$ column vector of real numbers.

Then, it is shown in such books are Roussas (1973) and the more advanced book by Muirhead (1982) that the characteristic function of a n-variate normal distribution has the form

$$\phi_n\left(\boldsymbol{u}\right) = E[\exp\left(i\boldsymbol{u}^T\boldsymbol{Z}\right)] = \exp\left(i\boldsymbol{u}^T\boldsymbol{\mu}_n - \frac{1}{2}\boldsymbol{u}_n^T\boldsymbol{\Gamma}_n\boldsymbol{u}\right). \tag{7.3.21}$$

When working with such a vector random variable we say it has a n-variate normal distribution with mean vector $\boldsymbol{\mu}_n$ and covariance matrix $\boldsymbol{\Gamma}_n$. In symbols, $\boldsymbol{Z} \sim \boldsymbol{N}\left(\boldsymbol{\mu}_n, \boldsymbol{\Gamma}_n\right)$.

Among the advantages of using the characteristic function to demonstrate that the family of normal distributions that characterize the Gaussian process under consideration is that multidimensional integrals, which arise in computing the marginal distributions of the process to demonstrate the family of distributions is consistent, may be represented in terms of their characteristic functions by substituting zeros in selected elements of the vector $\boldsymbol{u}$. By way of a simple illustration, suppose $n = 3$ so that the set of sites is $\mathbb{S} = (1, 2, 3)$. Then, given the characteristic function $\phi_3\left(\boldsymbol{u}\right)$, suppose it is desired to find the characteristic function corresponding to the subset $S_{12} = (1, 2)$. To derive this characteristic function. it suffices to substitute the vector $\boldsymbol{u}_{12}^T = (u_1, u_2, 0)$ into the formula for $\phi_3\left(\boldsymbol{u}\right)$ to obtain

$$\phi_{12}\left(\boldsymbol{u}_{12}\right) = \exp\left(i\boldsymbol{u}_{12}^T\boldsymbol{\mu}_{12} - \frac{1}{2}\boldsymbol{u}_{12}^T\boldsymbol{\Gamma}_{12}\boldsymbol{u}_{12}\right), \tag{7.3.22}$$

where

$$\boldsymbol{\mu}_{12} = \begin{pmatrix} \mu \\ \mu \end{pmatrix} \tag{7.3.23}$$

and

$$\boldsymbol{\Gamma}_{12} = \begin{pmatrix} \gamma\left(1,1\right) & \gamma\left(1,2\right) \\ \gamma\left(2,1\right) & \gamma\left(2,2\right) \end{pmatrix}. \tag{7.3.24}$$

From the form of the characteristic function $\phi_{12}\left(\boldsymbol{u}_{12}\right)$, it can be seen that the joint distribution of the random variables $Z\left(1\right)$ and $Z\left(2\right)$ is a bivariate normal with the indicated expectation vector $\boldsymbol{\mu}_{12}$ and covariance matrix

$\mathbf{\Gamma}_{12}$. Similarly, one could also derive the characteristic function for any other combination of two indices in the set $\mathbb{S} = (1, 2, 3)$ to demonstrate that these marginal distributions are also bivariate normals.

Furthermore, it can be shown by using the technique just outlined that the marginal distribution of the random variable $Z(1)$ would be the same if it were derived from either the joint distribution for the set $(1, 2)$ or that for the set $(1, 3)$ or from that whole set $(1, 2, 3)$. Moreover, similar remarks hold for the random variables $Z(2)$ and $Z(3)$. It is in this sense that for the case of $n = 3$ the family of normal distributions for the Gaussian process under consideration is consistent. From this illustrative example, it can also be seen that for any $n \geq 3$ and subsets $\mathbb{S}_k$ of $\mathbb{S}$ and any subset $\mathbb{S}_{k'}$ of $\mathbb{S}_k$ for $2 \leq k \leq n$ and $k' < k$, the characteristic function corresponding to the set $\mathbb{S}_{k'}$ would be the same no matter what set $\mathbb{S}_k$ it was derived from. Therefore, the family of normal distributions for the Gaussian process is consistent.

Before the study of mixing processes which are functions of Gaussian processes can proceed, it will be necessary to consider the problem of computing Monte Carlo realizations of the random $n \times 1$ vector $\boldsymbol{Z}$ for values of $n \geq 100$, where, in symbols,

$$\boldsymbol{Z} \sim \boldsymbol{N}(\boldsymbol{\mu}_n, \mathbf{\Gamma}_n). \tag{7.3.25}$$

More precisely, this problem can be stated as follows: Given specified numerical values of the parameters μ, σ and β as well as $n \geq 100$ so that the vector $\boldsymbol{\mu}_n$ and the covariance matrix $\mathbf{\Gamma}_n = (\gamma(s_1, s_2))$ can be computed, find one or more algorithms for computing Monte Carlo realizations of the random vector $\boldsymbol{Z}$. In the next section two solutions to this problem will be described, but the problem of choosing functions to map Monte Carlo realizations of the Gaussian process into intervals in $(0, 1)$ will be postponed to subsequent sections.

7.4 Computing Realizations of a Gaussian Process with Specified Covariance Function

In this section the practical problem of computing Monte Carlo realizations of a Gaussian process will be considered when the process has some specified covariance function $\gamma(s_1, s_2)$ defined for all pairs (s_1, s_2) of points in the set $\mathbb{S}$ of $n \geq 2$ nucleotide sites. In particular, the question to be addressed is for what values of n will it be practical to compute n realizations of the Gaussian process $Z(s)$ for all $s \in \mathbb{S}$ within an acceptable time span. As will be demonstrated, answers to this question will depend on the availability of software to compute the eigenvalues and eigenvectors the symmetric $n \times n$ covariance matrix $\mathbf{\Gamma}_n = (\gamma(s_1, s_2))$ determined by the covariance function as well as software to reduce this matrix to a lower triangular form.

With regard to factoring $\mathbf{\Gamma}_n$ into a product of lower triangular matrices one needs a result attributed to Choleski. According to the Choleski factorization theorem for any real non-singular positive definite matrix, see Kennedy and Gentle (1980), there exists a lower triangular matrix $\mathbf{L}_n$ such that

$$\mathbf{\Gamma}_n = \mathbf{L}_n \mathbf{L}_n^T \tag{7.4.1}$$

A procedure for computing $\mathbf{L}_n$ from the matrix $\mathbf{\Gamma}_n$ may also be found in the book by Kennedy and Gentle.

Turning to the problem of computing realizations of the $n \times 1$ vector $\boldsymbol{Z}$ for the Gaussian mixing process, let U_n be a $n \times 1$ vector of independent normal random variables with common mean 0 and variance 1. Then, as is well known, the vector random variable

$$\boldsymbol{Y}_n = \mathbf{L}_n \mathbf{U}_n \tag{7.4.2}$$

has a multivariate normal distribution with mean vector $E[\mathbf{Y}_n] = \mathbf{0}_n$ and covariance matrix

$$\begin{aligned} E\left[\mathbf{Y}_n \mathbf{Y}_n^T\right] &= E\left[\mathbf{L}_n \mathbf{U}_n \mathbf{U}_n^T \mathbf{L}_n^T\right] \\ &= \mathbf{L}_n E\left[\mathbf{U}_n \mathbf{U}_n^T\right] \mathbf{L}_n^T \\ &= \mathbf{L}_n \mathbf{I}_n \mathbf{L}_n^T = \mathbf{\Gamma}_n. \end{aligned} \tag{7.4.3}$$

The vector $\mathbf{Y}_n$, therefore, has the same covariance matrix as the vector $\mathbf{Z}$. To simulate a realization of the random vector $\mathbf{Z}$ one instructs the computer to compute the random vector

$$\mathbf{Z} = \boldsymbol{\mu}_n + \boldsymbol{Y}_n, \tag{7.4.4}$$

where $\boldsymbol{\mu}_n$ is a $n \times 1$ vector with each element equal to μ. As many software packages contain procedures for calculating the Choleski factorization, the procedure just outlined should work well for moderate values of n when the covariance matrix is non-singular.

However, when the random variables in the vector $\boldsymbol{Z}$ are highly correlated, the covariance matrix may be nearly singular and, therefore, other procedures should be used to simulate finitely many realizations of the vector random variable $\mathbf{Z}$. In such situations, it would be advisable to use a spectral decomposition of the vector $\boldsymbol{Y}_n$ by expressing it as a linear combination of the eigenvectors of .the matrix $\boldsymbol{\Gamma}_n$. Let λ_i be the i-th eigenvalue of $\boldsymbol{\Gamma}_n$ and let $\boldsymbol{\alpha}_i$ denote a $n \times 1$ eigenvector corresponding to λ_i for $i = 1, 2, \ldots, n$. It will be assumed that these eigenvectors form an orthonormal system. Next let W_k for $k = 1, 2, \ldots, n$ be scalar random variables such that

$$\boldsymbol{Y}_n = \sum_{k=1}^{n} W_k \boldsymbol{\alpha}_k. \tag{7.4.5}$$

To express W_k as a function of the vector $\boldsymbol{Y}_n$, multiply this equation on the left by $\boldsymbol{\alpha}_k^T$, the transpose of $\boldsymbol{\alpha}_k$. Because $\boldsymbol{\alpha}_k^T \boldsymbol{\alpha}_j = 0$ when $j \neq k$ and $\boldsymbol{\alpha}_k^T \boldsymbol{\alpha}_j = 1$ when $j = k$, it follows that

$$W_k = \boldsymbol{\alpha}_k^T \boldsymbol{Y}_n \tag{7.4.6}$$

for all $k = 1, 2, \ldots, n$.

As is well known and can be easily proved by using the characteristic function, every linear combination of a normal random vector $\boldsymbol{Y}_n$ with covariance matrix $\boldsymbol{\Gamma}_n$ has a scalar normal distribution with expectation

$$E\left[W_k\right] = E\left[\boldsymbol{\alpha}_k^T \boldsymbol{Y}_n\right] = 0 \tag{7.4.7}$$

and variance

$$\begin{aligned} var\left[W_k\right] = E\left[W_k^2\right] &= E\left[\boldsymbol{\alpha}_k^T \boldsymbol{Y}_n \boldsymbol{Y}_n^T \boldsymbol{\alpha}_k\right] \\ &= \boldsymbol{\alpha}_k^T \boldsymbol{\Gamma}_n \boldsymbol{\alpha}_k = \boldsymbol{\alpha}_k^T \boldsymbol{\alpha}_k \lambda_k = \lambda_k \end{aligned} \tag{7.4.8}$$

for all $k = 1, 2, \ldots, n$. Moreover, for $j \neq k$

$$cov\left[W_j W_k\right] = \boldsymbol{\alpha}_j^T \boldsymbol{\Gamma}_n \boldsymbol{\alpha}_k = 0. \tag{7.4.9}$$

Therefore, the random variables W_k are independent for $k = 1, 2, \ldots, n$.

Given these theoretical results, it is straight forward in principle to compute a realization of the random vector $\boldsymbol{Y}_n$. That is, compute realizations of the random variables W_k for $k = 1, 2, \ldots, n$, which are normally and independently distributed with a common expectation of 0 and variances λ_k for $k = 1, 2, \ldots, n$. Then compute a realization of the random vector $\boldsymbol{Y}_n$ as the sum

$$\boldsymbol{Y}_n = \sum_{k=1}^{n} W_k \boldsymbol{\alpha}_k. \tag{7.4.10}$$

When the covariance matrix $\boldsymbol{\Gamma}_n$ is nearly singular, many of its eigenvalues will be small and can, therefore, be neglected. In this case, let λ_i for

$i = 1, 2, \ldots, n_0$ denote the largest of the n eigenvalues. In some computer experiments that were conducted in applications of this theory with $n = 200$, n_0 was often less than 10. In such cases, it would suffice to use the sum

$$\boldsymbol{Y}_n = \sum_{k=1}^{n_0} W_k \boldsymbol{\alpha}_k \tag{7.4.11}$$

to compute a realization of the random vector $\boldsymbol{Y}_n$. Cases of this kind with nearly singular covariance matrices arose when the covariance function

$$\gamma\left(s_1, s_2\right) = \sigma^2 \exp\left(-\frac{\beta}{2}\left(s_1 - s_2\right)^2\right) \tag{7.4.12}$$

was used and β was chosen as a small value such as $\beta = 1/n^2$.

With respect to choosing values of n such that all eigenvalues and eigenvectors of the covariance matrix could be computed within reasonable time spans, a set of computer experiments were conducted with various choices of the parameters μ, σ and β and values of n. In each of these experiments, the Choleski. lower triangular matrix was also computed. In all these experiments, it was found that when $n \leq 500$ all computations could be done within a few seconds for values n of about 200, but, when $n = 500$ all computations could be done within a few minutes. For values of $n > 500$, however, the times taken to complete the calculations were larger and there was also evidence of numerical instabilities, particular when the covariance matrix was nearly singular. Even though the covariance function listed above has the advantage of providing a means for considering Gaussian mixing processes with highly correlated random variables, there appears to be a need for constructing Gaussian processes in such a way that computation with large covariance matrices can be avoided. In the next section, some classes of Gaussian processes will be considered such that realizations of the process can be computed without doing extensive computations with large covariance matrices.

7.5 Gaussian Processes That May be Computed Recursively

In this section, it will be shown that a Gaussian mixing process $Z\left(s\right)$ on the set $s \in \mathbb{S}$ of n sites can be formulated in terms of time series models that have been and are being used widely in statistics. Furthermore, it will be shown that realizations of this process may be computed recursively so that problems that may arise in connection with carrying out numerical operations on large covariance matrices can be avoided. As in previous sections, it will be assumed that $E\left[Z\left(s\right)\right] = \mu$, a constant, for all $s \in \mathbb{S}$. It will thus be sufficient to consider a process $Y\left(s\right)$ defined by $Y\left(s\right) = Z\left(s\right) - \mu$ for all $s \in \mathbb{S}$.

One of the simplest cases for which it is possible to compute realizations of the Y-process recursively is when it satisfies the stochastic difference equation

$$Y(s) = \beta Y(s-1) + \epsilon(s) \tag{7.5.1}$$

for $s \geq 2$, where β is a constant parameter and $\epsilon' s$ are normal independent random variables with a common expectation of 0 and variance σ^2. In the time series literature, this random difference equation is also known as an auto-regressive model of order one. All solutions of this equation depend on two parameters, β and σ^2, and it is natural to ask what condition must the parameter β satisfy in order that the solution is a stationary Gaussian process. When analyzing this equation, it will simplify the presentation if it is supposed that this stochastic difference equation is defined on the set

$$\mathbb{N} = (s \mid s = 0, \pm 1, \pm 2, \ldots) \tag{7.5.2}$$

of all integers, but in when doing computations attention will be confined to the finite set $\mathbb{S}$ of sites.

By iterating this equation, it can be shown that

$$Y(s) - \sum_{\nu=0}^{k} \beta^{\nu} \epsilon(s-\nu) = \beta^{k+1} Y(s-(n+1)) \tag{7.5.3}$$

for all $s \in \mathbb{N}$, and, by squaring and taking expectations, it follows that

$$\begin{aligned} &E\left[\left(Y(s) - \sum_{\nu=0}^{k} \beta^{\nu} \epsilon(s-\nu)\right)^2\right] \\ &= \beta^{2(k+1)} E\left[Y^2(s-(n+1))\right]. \end{aligned} \tag{7.5.4}$$

Letting $k \uparrow \infty$ in the sum of the left suggests that the solution of the stochastic difference equation has the form

$$Y(t) = \sum_{\nu=0}^{\infty} \beta^{\nu} \epsilon(t-\nu) \tag{7.5.5}$$

for all $s \in \ \mathbb{N}$, where the random infinite series, which is also known as an infinite moving average, must converge in some sense. In order that the

Y-process be stationary and Gaussian, it is necessary that the expectation $E\left[Y^2(s)\right]$ be finite for all s. Because the $\epsilon's$ are not correlated, it can be seen by squaring and formally taking expectations in the above equation that

$$\begin{aligned} E\left[Y^2(s)\right] &= \sigma^2 \sum_{v=0}^{\infty} \beta^{2v} \\ &= \frac{\sigma^2}{1-\beta^2} \end{aligned} \tag{7.5.6}$$

if, and only if, $|\beta| < 1$. Moreover, when this condition is satisfied, the right-hand side (7.5.4) converges to 0 as $k \to \infty$ and the random infinite series is said to converge in quadratic mean to a solution of the auto-regressive equation.

So far no mention has been made as to whether it is mathematically valid to take expectations term by term in an infinite random series so that the resulting infinite series converges to a valid formula. However, it is well known that for the case of convergence in quadratic mean, the operations of taking expectations and infinite sums can be interchanged with impunity. Thus, for $h \geq 0$, the auto-covariance function of the Y-process is given by

$$\begin{aligned} \gamma(s, s+h) &= E\left[Y(s)Y(s+h)\right] \\ &= E\left[\sum_{v_1=0}^{\infty}\sum_{v_2=0}^{\infty} \beta^{v_1+v_2}\epsilon(s-v_1)\epsilon(s+h-v_2)\right] \\ &= \sigma^2\beta^h \sum_{\nu=0}^{\infty} \beta^{2\nu} = \frac{\sigma^2\beta^h}{1-\beta^2}. \end{aligned} \tag{7.5.7}$$

Observe that all expectations in the double sum on the right are 0 except when $s - \nu_1 = s + h - \nu_2$, which implies that $\nu_2 = \nu_1 + h$ and

$$E\left[\epsilon(s-v_1)\epsilon(s+h-v_2)\right] = \sigma^2. \tag{7.5.8}$$

From this result, it is easy to see that for $h \geq 0$ the auto-correlation function of the process is

$$\rho(h) = \frac{\gamma(s, s+h)}{\gamma(s,s)} = \beta^h. \tag{7.5.9}$$

If $0 < \beta < 1$, then this function is positive for all $h > 0$, but if $-1 < \beta < 0$, then $\rho(h)$ is positive or negative, depending on whether h is an even or odd positive integer. From a purely theoretical point of view, it may be some interest to consider a Gaussian mixing process with this auto-correlation function for $-1 < \beta < 0$ in the class of nucleotide substitution models under consideration. In passing, it should be mentioned that a stationary Gaussian process generated by a first order auto-regressive model has the Markov property.

Given these results, it is easy to see how realizations of the Y-process on the set $\mathbb{S}$ of nucleotide sites may be computed recursively. Suppose, the elements of $\mathbb{S}$ are ordered from 1 to n. To set up a recursive procedure, compute a realization of each of the random variable $\epsilon\left(s\right)$ for $s = 1, 2, \ldots, n$ by calling n independent normal random variables with a common expectation of 0 and variance σ^2. Next, compute one realization of the random variable $Y\left(1\right)$ with a normal distribution expectation 0 and variance $\sigma^2/\left(1 - \beta^2\right)$. Then, a realization of the random variable $Y\left(2\right)$ would be computed using the equation

$$Y\left(2\right) = \beta Y\left(1\right) + \epsilon\left(2\right). \tag{7.5.10}$$

To proceed recursively, if a realization $Y\left(k - 1\right)$ has been computed following this recursive scheme, then a realization of the random variable $Y\left(k\right)$ would be computed, using the equation

$$Y\left(k\right) = \beta Y\left(k - 1\right) + \epsilon\left(k\right) \tag{7.5.11}$$

for $k = 2, 3, \ldots, n$. Finally, to compute realizations of the Z-process, set $Z\left(k\right) = \mu + Y\left(k\right)$ for $k = 1, 2, \ldots, n$.

It is of interest to extend a first order auto-regression to one of second order. A second order auto-regressive model of the form

$$Y(s) = \beta_1 Y(s-1) + \beta_2 Y(s-2) + \epsilon(s) \tag{7.5.12}$$

is a straight-forward extension of the first order process, where β_1 and β_2 are parameters, the $\epsilon' s$ are independent Gaussian random variables with common expectation 0 and variance σ^2, and $s \in \mathbb{N}$. For this model, one may ask for conditions on the β-parameters such that there exists a sequence (δ_v) with the property that the infinite series

$$\sum_{v=0}^{\infty} \delta_v^2 \tag{7.5.13}$$

converges and a solution of the second order stochastic difference equation has the form of an infinite moving average

$$Y(t) = \sum_{v=0}^{\infty} \delta_v \epsilon(t - v). \tag{7.5.14}$$

It can be shown that the convergence of the series in (7.5.13) implies the random series in (7.5.14) converges in quadratic mean. In order for (7.5.13) to converge, it suffices to require that the series

$$\sum_{v=0}^{\infty} |\delta_v| \tag{7.5.15}$$

converge. For, if this series converges, then $|\delta_n| \rightarrow 0$ and $n \rightarrow \infty$ and there is an n_0 such that $n \geq n_0$ implies $|\delta_n|^2 \leq |\delta_n|$. From (7.5.14 it can be seen that

$$E\left[\epsilon(t-v)Y(t)\right] = \delta_v \sigma^2 \tag{7.5.16}$$

for all $v \geq 0$ and $t \in \mathbb{N}$. Therefore, by multiplying equation second order auto-regression equation by $\epsilon(t-v)Y(t)$ and taking expectations it can be seen that the sequence (δ_v) must satisfy the second order difference equation

$$\delta_v = \beta_1 \delta_{v-1} + \beta_2 \delta_{v-2}. \tag{7.5.17}$$

We seek a solution of this equation such that $\delta_v = 0$ if $v < 0$. Under this condition, if δ_0 and δ_1 are specified, then the sequence (δ_v) may be determined recursively for $v \geq 2$, but it will be necessary to find solutions such that the resulting infinite series converges. From now on, let $\delta_0 = 1$ and from this assignment it can be seen that $\delta_1 = \beta_1$, since $\delta_{-1} = 0$.

Suppose one searches for solutions of the second order difference equation of the form $\delta_v = r^v$, where r is a constant. Then, it can be shown that r is a root of the quadratic equation

$$x^2 - \beta_1 x - \beta_2 = 0. \tag{7.5.18}$$

The roots of this equation are

$$\begin{gathered} r_1 = \frac{1}{2}\beta_1 + \frac{1}{2}\sqrt{\left(\beta_1^2 + 4\beta_2\right)} \\ \textit{and} \\ r_2 = \frac{1}{2}\beta_1 - \frac{1}{2}\sqrt{\left(\beta_1^2 + 4\beta_2\right)}, \end{gathered} \tag{7.5.19}$$

which may be complex.

If these roots are distinct, then a solution of the equation may be represented in the form

$$\delta_v = c_1 r_1^v + c_2 r_2^v, \tag{7.5.20}$$

where the constants c_1 and c_2 are the solution of the equations

$$\begin{aligned} 1 &= c_1 + c_2 \\ \beta_1 &= c_1 r_1 + c_2 r_2. \end{aligned} \tag{7.5.21}$$

Therefore, if the roots r_1 and r_1 lie in the unit circle, i.e., $\lfloor r_1 \rfloor < 1$ and $\lfloor r_2 \rfloor < 1$, the infinite series in (7.5.15) will converge. If $r_1 = r_2 = r$, then it can be shown that this series will also be convergent if $\lfloor r \rfloor < 1$. When numerically specifying a model for computer experiments, it may be of interest to specify the roots r_1 and r_2 such that they lie in the unit circle. Then,

$$(x - r_1)(x - r_2) = x^2 - (r_1 + r_2)x + r_1 r_2, \tag{7.5.22}$$

which yields the values $\beta_1 = r_1 + r_2$ and $\beta_2 = -r_1 r_2$ for the β-parameters.

When the roots r_1 and r_2 lie in the unit circle, the auto-covariance function of the process is given by the convergent series

$$\gamma(s, s+h) = \sum_{v=0}^{\infty} \delta_v \delta_{v+h} \sigma^2 \tag{7.5.23}$$

for $h \geq 0$ and all $s \in \mathbb{N}$. But, this may not be a desirable form of the function if one wishes to compute its values. From now on, let $\gamma(s, s+h) = \gamma(h) = \gamma(-h)$. By observing that

$$\gamma(h) = E\left[Y(s-h)Y(s)\right] \tag{7.5.24}$$

for all $h \geq 0$ and s, it can be shown that this function also satisfies a second order difference equation such that

$$\begin{aligned}
\gamma(0) &= \beta_1 \gamma(1) + \beta_2 \gamma(2) + \sigma^2, \\
\gamma(1) &= \beta_1 \gamma(0) + \beta_2 \gamma(1), \\
&\quad \textit{and} \\
\gamma(h) &= \beta_1 \gamma(h-1) + \beta_2 \gamma(h-1)
\end{aligned} \tag{7.5.25}$$

for $h \geq 2$.

For $h = 2$ this is a system in three unknowns and a call to symbolic computation engine yields the symbolic solution

$$\begin{aligned}
&\gamma(0) \\
&= (-1 + \beta_2) \frac{\sigma^2}{(1 + \beta_2)(\beta_1 - 1 + \beta_2)(\beta_1 + 1 - \beta_2)}, \\
&\gamma(1) \\
&= -\beta_1 \frac{\sigma^2}{(1 + \beta_2)(\beta_1 - 1 + \beta_2)(\beta_1 + 1 - \beta_2)}, \\
&\textit{and} \\
&\gamma(2) = -\sigma^2 \frac{\beta_1^2 + \beta_2 - \beta_2^2}{(1 + \beta_2)(\beta_1 - 1 + \beta_2)(\beta_1 + 1 - \beta_2)}.
\end{aligned} \tag{7.5.26}$$

From an inspection of this symbolic system, it can be seen that one must exclude the case $\beta_2 = -1$ to ensure all the above formulas yield finite numbers. Similarly, to ensure that the process has a non-zero variance, i.e. $\gamma(0) \neq 0$, the case $\beta_2 = 1$ must be excluded. Given numerical values of $\gamma(0)$ and $\gamma(1)$, values of $\gamma(h)$ may be computed recursively for $h \geq 0$. It is interesting to observe, that if the roots r_1 and r_2 are specified numerically with values in the unit circle, then it would be possible to compute variables of the auto-covariance function $\gamma(h)$ recursively.

To compute realizations of the Y-process in this case, it will be necessary to initialize the recursion procedure by computing realizations of the random variables $Y(1)$ and $Y(2)$, which have a bivariate normal distribution with expectation vector $\mathbf{0}$ and the covariance matrix

$$\begin{pmatrix} \gamma(0) & \gamma(1) \\ \gamma(1) & \gamma(0) \end{pmatrix}. \tag{7.5.27}$$

In general, if two random variables X_1 and X_2 have a bivariate normal distribution with expectation vector

$$\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \tag{7.5.28}$$

and covariance matrix

$$\begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix}, \tag{7.5.29}$$

then $X_1 \sim N\left(\mu_1, \sigma_1^2\right)$ and the conditional distribution of X_2, given $X_1 = x_1$, is normal with expectation

$$E\left[X_2 \mid X_1 = x_1\right] = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}\left(x_1 - \mu_1\right) \tag{7.5.30}$$

and variance $\sigma_2^2\left(1 - \rho^2\right)$. A derivation of these formulas may be found in almost any book on mathematical statistics. For the special case under consideration, $\mu_1 = \mu_2 = 0$, $\sigma_1^2 = \sigma_2^2 = \gamma(0)$ and $\rho = \gamma(1)/\gamma(0)$. To approximate values of $\gamma(0)$ and $\gamma(1)$, one could use partial sums of the series

$$\gamma(h) = \sum_{v=0}^{\infty} \delta_v \delta_{v+h} \sigma^2 \tag{7.5.31}$$

for $h = 0$ and $h = 1$.

Therefore, to compute a realization of the random variable $Y(1)$ draw a sample of size 1 from a normal distribution with expectation 0 and variance $\gamma(0)$. Then, given $Y(1) = y$, to simulate a value of $Y(2)$ draw a sample of size 1 from a normal distribution with expectation ρy and variance $\gamma(0)\left(1-\rho^2\right)$. Given the values $Y(1)$ and $Y(2)$, realizations of $Y(k)$ for $k \geq 3$ may be computed recursively from the equation

$$Y(k) = \beta_1 Y(k-1) + \beta_2 Y(k-2) + \epsilon(k) \tag{7.5.32}$$

for $k = 3, 4, \ldots, n$. In this recursive equation realizations of the independent random variables $\epsilon(k)$ for $k = 1, 2, \ldots, n$ would be computed in advance. Each of these random variables is normally distributed with an expectation of 0 and variance σ^2. Finally, to compute values of the Z-process, let $Z(k) = \mu + Y(k)$ for $k = 1, 2, \ldots, n$.

The two illustrative examples just described, based on first and second order auto-regressive models, may be generalized in countless ways and belong to a vast literature on time series. For example, a model of the form

$$Y(t) = \beta Y(t-1) + \epsilon(t) + \alpha\epsilon(t-1) \tag{7.5.33}$$

is known as first order auto-regressive, moving average process, where α and β are constant parameters and the $\epsilon' s$ may be Gaussian noise. Books on the subject include those of Brillinger (1981), Brockwell and Davis (1991), and Fuller (1976). Stochastic difference equations had been treated in the literature on stochastic processes for several decades, but it was not until the book by Box and Jenkins (1976) was published that variations of auto-regressive models were widely used in not only analyzing data on time series but also in attempts to deduce a model that may have generated the data.

7.6 Monte Carlo Implementation of Mixtures of Transition Rates for Markov Processes

When the number of nucleotide sites n is large, it becomes very difficult to analyze the mixture of Markov nucleotide substitution processes described in the previous sections. However, as will be shown in this section, the algorithms used to simulate Monte Carlo realizations of the process can be described and implemented in a straight forward manner. The first step in defining these algorithms is to find a procedure for computing realizations of mutation rate random variables $\Theta(s)$ for $s = 1, 2, \ldots, n$, which take values in the interval $(0, 1)$ as functions of the Gaussian mixing process $Z(s)$ for $s = 1, 2, \ldots, n$. Formally, we need to find some function $h(z)$ for $z \in \mathbb{R}$ that may be computed with ease such that $\Theta(s) = h(Z(s)) \in (0, 1)$ for all $s = 1, 2, \ldots, n$.

To this end, let X denote a random variable taking values in some subset (interval) D of $\mathbb{R}$ and let

$$F\left(x\right)=P\left[X\leq x\right] \tag{7.6.1}$$

for $x\in D$ denote its distribution function. The function $F\left(x\right)$ is non-decreasing on D and takes values in the interval $(0,1)$. That is for any two points x_1 and x_2 in D such that if $x_1\leq x_2$, then $y_1=F\left(x_1\right)\leq F\left(x_2\right)=y_2$. For the purposes that follow, it will also be assumed that $F\left(x\right)$ is strictly increasing on D so that if $x_1<x_2$, then $y_1=F\left(x_1\right)<F\left(x_2\right)=y_2$. For such functions, for every $x\in D$, there is a unique $y\in(0,1)$ such that $y=F\left(x\right)$ and there exists an inverse function $F^{(-1)}\left(y\right)$ defined for all $y\in(0,1)$ and taking values in D such that if $y=F\left(x\right)$, then $F^{(-1)}\left(y\right)=x$. Furthermore, this inverse function is strictly increasing. For suppose $y_1<y_2$, then there is a $x_1<x_2$ such that $y_1=F\left(x_1\right)$ and $y_2=F\left(x_2\right)$. Therefore, $F^{(-1)}\left(y_1\right)=x_1<x_2=F^{(-1)}\left(y_2\right)$.

Now consider a random variable Y taking values in $(0,1)$ defined by $Y=F\left(X\right)$ with distribution function

$$F_Y\left(y\right)=P\left[Y\leq y\right] \tag{7.6.2}$$

for $y\in(0,1)$. Then, because $F^{(-1)}\left(y\right)$ is an increasing function on $(0,1)$, it follows that

$$\begin{aligned}F_Y\left(y\right)&=P\left[Y\leq y\right]=P\left[F\left(X\right)\leq y\right]\\&=P\left[X\leq F^{(-1)}\left(y\right)\right]=F\left(F^{(-1)}\left(y\right)\right)\\&=F\left(x\right)=y\end{aligned} \tag{7.6.3}$$

for all $y\in(0,1)$. The next step is to identify the distribution function $F_Y\left(y\right)$.

A random variable U is said to have a uniform distribution on the interval $(0,1)$ if its density function is

$$f\left(u\right)=1 \tag{7.6.4}$$

for all $u\in(0,1)$ and $f\left(u\right)=0$ for all $u\notin(0,1)$. Therefore, the distribution function of U is

$$F_U\left(u\right)=\int_0^u f\left(s\right)ds=u \tag{7.6.5}$$

for all $u\in(0,1)$, which is the same as the distribution function of the random variable $Y=F\left(X\right)$ defined above. We have thus proved a rather remarkable result which states that for any strictly increasing distribution function $F(x)$ of a random variable X defined for all $x \in D$, some subset of $\mathbb{R}$, the .random variable $Y = F(X)$ has a uniform distribution on the interval $(0,1)$.

This result also suggests that if the random variable U has a uniform distribution of the interval $(0,1)$, then solving the equation

$$F(X) = U \tag{7.6.6}$$

.for X would provide a formula for simulating realizations of a random variable X with distribution function $F(x)$. One is thus led to consider a random variable W defined by

$$W = F^{(-1)}(U). \tag{7.6.7}$$

To prove that the random variable W is equal in distribution to the random variable X, it must be shown that the distribution function of the random variable W is $F(x)$ for all $x \in D$. By definition, the distribution function of the random variable W is

$$\begin{aligned} F_W(w) &= P[W \leq w] = P\left[F^{(-1)}(U) \leq w\right] \\ &= P[U \leq F(w)] = F(w) \end{aligned} \tag{7.6.8}$$

for all $w \in D$. Therefore, the random variables W and X are equal in distribution.

This method of simulating realizations of random variables is particularly useful when a random variable X has an exponential distribution on the interval $D = [0,\infty) = (x \in \mathbb{R}\, |0 \leq x < \infty)$ with density function

$$f(x) = \beta \exp(-\beta x) \tag{7.6.9}$$

for all $x \in D$, where $\beta > 0$ is a positive parameter. For this case, the distribution function of X is

$$F(x) = \int_0^x f(s)\, ds = 1 - \exp(-\beta x) \tag{7.6.10}$$

for all $x \in D$ so that we are led to solve the equation

$$1 - \exp(-\beta X) = U \tag{7.6.11}$$

for X, where U is a uniform random variable on $(0,1)$. It is easy to see that

$$X = -\frac{1}{\beta} \ln(1-U). \tag{7.6.12}$$

One could apply this formula to simulate realizations of an exponential random variable X, but when the number of realizations of X is a large number, it is natural to ask whether the subtraction $1-U$ could be avoided.

Let V denote the random variable $1-U$. Then, the distribution function of V for all $v \in (0,1)$ is

$$\begin{aligned} F_V(v) &= P[V \leq v] = P[1-U \leq v] \\ &= P[U \geq 1-v] = 1-(1-v) = v. \end{aligned} \tag{7.6.13}$$

Therefore, the random variable $V = 1-U$ and U are equal in distribution. Consequently, when computing a large number of realizations of a random variable X with an exponential distribution depending on a parameter $\beta > 0$, it suffices to use the formula

$$X = -\frac{1}{\beta}\ln(U). \tag{7.6.14}$$

Observe that this formula maps the interval $(0,1)$ into the interval $D = [0,\infty)$. In what follows, this formula will be used extensively when computing realizations of sojourn times in states for the mixtures of Markov nucleotide substitution processes under consideration.

Having described some fundamental principles underlying Monte Carlo simulations procedures, the next topic to consider is that of computing realizations of the random variable $\Theta(s)$ for $s = 1,2,\ldots,n$ mentioned at the start of this section. A random variable Z has a normal distribution with expectation 0 and variance 1 (in symbols $Z \sim N(0,1)$), if its density is

$$f(z) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right) \tag{7.6.15}$$

for all $z \in \mathbb{R}$. The distribution function of Z is

$$P[Z \leq z] = \Phi(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}\exp\left(-\frac{s^2}{2}\right)ds \tag{7.6.16}$$

for all $z \in \mathbb{R}$. By construction, the random variables in the Gaussian mixing process $Z(s)$ for $s = 1,2,\ldots,n$ each have a normal distribution with expectation μ and variance σ^2. In symbols, $Z(s) \sim N(\mu,\sigma^2)$ for all $s = 1,2,\ldots,n$. For every s, the distribution function of $Z(s)$ is

$$\begin{aligned} F(z) &= P[Z(s) \leq z] = P\left[\frac{Z(s)-\mu}{\sigma} \leq \frac{z-\mu}{\sigma}\right] \\ &= P\left[Z \leq \frac{z-\mu}{\sigma}\right] = \Phi\left(\frac{z-\mu}{\sigma}\right) \end{aligned} \tag{7.6.17}$$

for all $z \in \mathbb{R}$.

Therefore, by a result demonstrated above, it follows that the random variable

$$U(s) = \Phi\left(\frac{Z(s)-\mu}{\sigma}\right) \tag{7.6.18}$$

has a uniform distribution on the interval $(0, 1)$ for all $s = 1, 2, \ldots, n$. Given this collection of uniform random variables, it will be possible to map them to chosen subintervals of $(0, 1)$ when computing realizations of rate matrices for Markov nucleotide substitution processes at some site on a molecule of DNA. Suppose, for example, that the nucleotide at site s is i and one wishes to compute a realization of the rate $\theta_{ij}(s)$ for the nucleotide substitution $i \to j$, which may depend on whether the substitution is a transition of transversion. It is thought by many investigators that the transition rates are higher than transversions rates.

To accommodates such ideas, by way of an illustration, suppose all transition rates lie in the interval $\left[10^{-5}, 10^{-4}\right]$ and all transversion rates lie the interval $\left[10^{-6}.10^{-5}\right]$. Then, the linear function

$$Y_1(s) = 10^{-5} + \left(10^{-4} - 10^{-5}\right) U(s) \tag{7.6.19}$$

maps the interval $(0, 1)$ into the interval $\left[10^{-5}, 10^{-4}\right]$, and similarly, the linear function

$$Y_2(s) = 10^{-6} + \left(10^{-5} - 10^{-6}\right) U(s) \tag{7.6.20}$$

maps the interval $(0, 1)$ into the interval $\left[10^{-6}.10^{-5}\right]$. Then, if the substitution $i \to j$ is a transition, let $\theta_{ij}(s) = Y_1(s)$, and if it is a transversion, let $\theta_{ij}(s) = Y_2(s)$. Observe that for any site s at most three substitution (mutation) rates will need to be computed. Also observe that the parameter in the exponential distribution, governing the sojourn time in sites s when i is the nucleotide at this site, is

$$\sum_{j \neq i} \theta_{ij}(s). \tag{7.6.21}$$

To implement the ideas just described, it will be necessary to have an efficient method for computing values of the normal distribution function

$$\Phi\left(\frac{Z(s) - \mu}{\sigma}\right) \tag{7.6.22}$$

for all $s = 1, 2, \ldots, n$. As is well known, this distribution function cannot be expressed in terms of the elementary functions from the calculus, but, fortunately good approximations to this function, which can be computed efficiently, have been developed. The book, Kennedy and Gentle (1980) may be consulted for a description of several methods for obtaining good approximation to the distribution function $\Phi(z)$. Further discussion of these methods of approximation may also be found in the book Thisted (1988).

Interestingly, the collection of random variables $U(s)$ for $s = 1, 2, \dots, n$ are uniformly distributed on the interval $(0, 1)$ but they are not independent. To see this, let $\mathbf{\Gamma}_k$ denote the covariance matrix for any subset $\mathbb{S}_k$ of $\mathbb{S}$ with k sites for $2 \leq k \leq n$ for the Gaussian mixing process, and let

$$\boldsymbol{Z}_k = (Z(s) \mid s \in \mathbb{S}_k) \tag{7.6.23}$$

denote a $k \times 1$ vector of Gaussian random variables with corresponding expectation vector $\boldsymbol{\mu}_k$. Then, as is well known, if the matrix $\mathbf{\Gamma}_k$ is nonsingular, the density of the random vector $\boldsymbol{Z}_k$ is

$$f_k(\boldsymbol{z}_k) = (2\pi)^{-\frac{k}{2}} (\det \mathbf{\Gamma}_k)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left((\boldsymbol{z}_k - \boldsymbol{\mu}_k)^T \mathbf{\Gamma}_k^{-1} (\boldsymbol{z}_k - \boldsymbol{\mu}_k)\right)\right), \tag{7.6.24}$$

where $\det \mathbf{\Gamma}_k$ is the determinant of $\mathbf{\Gamma}_k$. Therefore, for any two sites s_1 and s_2 consider the expectation

$$E[U(s_1) U(s_2)] = \int_{\mathbb{R}_2} \Phi\left(\frac{z_1 - \mu}{\sigma}\right) \Phi\left(\frac{z_2 - \mu}{\sigma}\right) f_2(\boldsymbol{z}_2)\, d\boldsymbol{z}_2. \tag{7.6.25}$$

In general,

$$E[U(s_1) U(s_2)] \neq E[U(s_1)] E[U(s_2)] \tag{7.6.26}$$

so that the random variables $U(s_1)$ and $U(s_2)$ will not be independent. In principle, the expectation $E[U(s_1) U(s_2)]$ may be expressed as function of the parameters μ, σ and the covariance function $\gamma(s_1, s_2)$ but the expression of this function will not be pursued here.

The following material will be a digression from the main theme of this section, but will be included because of its general interest for the construction set of random variable that are not independent. Among the problems that arise in designing Monte Carlo simulation models that are often used in biology is that of developing procedures for simulating realizations of a sequence random variables taking values in the interval $[0, \infty)$. For example, one may wish to compute realizations of exponential random variables with a common parameter $\beta > 0$ that are not independent. In applications, such random variables often have the interpretation of being waiting times among events that are not independent. By way of an illustration, for the Gaussian mixing process under consideration, one could compute these realizations of these waiting times by using the formula

$$X(s) = -\frac{1}{\beta} \ln(U(s)) = -\frac{1}{\beta} \ln(\Phi\left(\frac{Z(s) - \mu}{\sigma}\right). \tag{7.6.27}$$

for $s = 1, 2, \ldots, n$. It is interesting to note that this collection of exponential random variables has a common exponential distribution with parameter $\beta > 0$ but they are not distributed independently.

When one wishes to consider random variables as waiting times among events, there are other interesting approaches for mapping normal random variables taking values in the set $\mathbb{R} = (-\infty.\infty)$ of real numbers to the set $[0, \infty)$ of non-negative real numbers. An example of such a mapping is the function

$$h(x) = \exp(x) \tag{7.6.28}$$

defined for all $x \in \mathbb{R}$. Thus, among the choices for waiting time random variables are those defined by

$$Y(s) = \exp(Z(s)) \tag{7.6.29}$$

for all $s \in \mathbb{S}$. For every s, the random variable $Z(s)$ has a normal distribution with expectation μ and variance σ^2. In symbols, $Z(s) \sim N\left(\mu, \sigma^2\right)$. Therefore, the expectation of $Y(s)$ is the constant

$$E[Y(s)] = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp(y) \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) dy = \exp\left(\mu + \frac{\sigma^2}{2}\right) \tag{7.6.30}$$

for all $s \in \mathbb{S}$. Note that this integral is easy to evaluate, because it is the moment generating function of a normal distribution evaluated at 1. The process just described is called a stationary log-Gaussian process and in principle it would be possible to work out a formula of the covariance function of the process, but this exercise will be left to the reader.

7.7 Transition Rates Based on Logistic Gaussian Processes

In the preceding section, transition rate random variables $\Theta(s)$ for $s = 1, 2, \ldots, n$ were introduced as functions of a Gaussian mixing process $Z(s)$ with the property that they were uniformly distributed on some sub-intervals of $(0, 1)$, but it would also be of interest to consider transition rate random variables that would have non-uniform distributions on these sub-intervals. Furthermore, it would be helpful if the properties of the transition rate distributions could be easily analyzed in terms of the parameters of the Gaussian mixing process. The problem of finding such a distribution reduces to that of making a judicious choice of a function $h(x)$ on $\mathbb{R}$ such that $Y(s) = h(Z(s)) \in (0, 1)$ for all $s = 1, 2, \ldots, n$.

Among the judicious choices of the function $h(x)$ is the distribution function of logistic distribution which has the simple formula

$$y = H(x) = \frac{\exp(x)}{1+\exp(x)} \tag{7.7.1}$$

defined for all $x \in \mathbb{R}$. To begin investigating the properties of this function, which maps $\mathbb{R}$ into $(0,1)$, it will be helpful to derive its inverse. To this end, consider the ratio

$$\frac{y}{1-y} = \exp(x). \tag{7.7.2}$$

From this ratio, it can be seen that for $y \in (0,1)$, the inverse function is

$$H^{(-1)}(y) = \ln\left(\frac{y}{1-y}\right) = x. \tag{7.7.3}$$

One is thus led to consider a random variable Y taking values in $(0,1)$ defined by

$$Y = H(X) = \frac{\exp(X)}{1+\exp(X)}, \tag{7.7.4}$$

where $X \sim N\left(\mu, \sigma^2\right)$. By definition, the distribution function of Y for $y \in (0,1)$ is

$$\begin{aligned} G(y) &= P[Y \le y] = P[H(X) \le y] \\ &= P\left[X \le H^{(-1)}(y)\right] = P\left[X \le \ln\left(\frac{y}{1-y}\right)\right] \\ &= P\left[\frac{X-\mu}{\sigma} \le \frac{\ln\left(\frac{y}{1-y}\right)-\mu}{\sigma}\right] = \Phi\left(\frac{\ln\left(\frac{y}{1-y}\right)-\mu}{\sigma}\right), \end{aligned} \tag{7.7.5}$$

where $\Phi(z)$ is the distribution function of a normal random variable Z such that $Z \sim N(0,1)$. Therefore, the density function of the random variable Y is

$$\begin{aligned} g(y) &= \frac{d}{dy}\Phi\left(\frac{\ln\left(\frac{y}{1-y}\right)-\mu}{\sigma}\right) \\ &= \frac{1}{\sqrt{2\pi}\sigma y(1-y)} \exp\left(-\frac{1}{2\sigma^2}\left(\frac{\ln\left(\frac{y}{1-y}\right)-\mu}{\sigma}\right)\right) \end{aligned} \tag{7.7.6}$$

for all $y \in (0,1)$. In what follows, the distribution of the random variable Y will be referred to as the logistic-normal with parameters μ and σ.

It is very difficult to derive closed formulas for the expectation and variance of the random variable Y with logistic normal distribution. Fortunately, however, it is easy to derive a simple formula for the quantiles of the logistic normal distribution. For any $q \in (0,1)$ the q-th quantile of the logistic normal distribution is a number $y_q \in (0,1)$ such that

$$q = G\left(y_q\right) = \Phi\left(\frac{\ln\left(\frac{y_q}{1-y_q}\right) - \mu}{\sigma}\right) = \Phi\left(z_q\right), \tag{7.7.7}$$

where z_q is the q-th quantile of a random variable $Z \sim N(0,1)$. Therefore,

$$\frac{\ln\left(\frac{y_q}{1-y_q}\right) - \mu}{\sigma} = z_q, \tag{7.7.8}$$

which implies that

$$y_q = \frac{\exp\left(\mu + \sigma z_q\right)}{1 + \exp\left(\mu + \sigma z_q\right)}. \tag{7.7.9}$$

Given a value q, one may use the inverse function $\Phi^{(-1)}(y)$ of $\Phi(z)$ to find z_q, the q-th quantile of Z by using the equation $z_q = \Phi^{(-1)}(q)$. In the book, Kennedy and Gentle (1980), several methods for approximating this inverse function are described. Furthermore, most statistical software packages contain programs for approximating this inverse function as well as the distribution function $\Phi(z)$.

As is well known, the median of the standard normal distribution is $z_{0.5} = 0$, which implies that the median of the logistic normal distribution is

$$y_{0.5} = \frac{\exp(\mu)}{1 + \exp(\mu)} \tag{7.7.10}$$

for any value of $\mu \in \mathbb{R}$. In particular, if $\mu = 0$, then the median of the logistic normal distribution is $y_{0.5} = 0.5$. For the case $\mu = 0$, it is also of interest to describe a procedure for choosing another quantile of the logistic normal distribution to determine a value of σ. For example, suppose one wished to choose a value of σ such that $y_q = 0.75$ for $q = 0.9750$. To choose this value, it is helpful to use the inverse equation

$$\ln\left(\frac{y_q}{1-y_q}\right) = \sigma z_q, \tag{7.7.11}$$

which yields the formula

$$\sigma = \frac{1}{z_q} \ln\left(\frac{y_q}{1-y_q}\right) \tag{7.7.12}$$

for σ. For $q = 0.9750$ and $z_q = 1.96$, and a call to the computation engine yields the value

$$\sigma = \frac{1}{1.96} \ln \left(\frac{0.75}{1 - 0.75} \right) = 0.560\,516\,473\,810\,26. \tag{7.7.13}$$

Given this value of σ, one could use the formula

$$y_q = \frac{\exp(-\sigma \times 1.96)}{1 + \exp(-\sigma \times 1.96)} \tag{7.7.14}$$

to find that $y_q = 0.25$ for $q = 0.0250$. One could, therefore, conclude that if the random variable Y has a logistic normal distribution with parameters $\mu = 0$ and $\sigma = 0.560\,516\,473\,810\,26$, the probability that Y lies in the interval $(0.25, 0.75)$ is

$$P[0.25 \leq Y \leq 0.75] = \Phi(1.96) - \Phi(-1.96) = 0.95. \tag{7.7.15}$$

Interestingly, this exercise could also been carried out for values of μ other than $\mu = 0$ but such exercises will be left to the reader. To choose a median for the random variable Y, for example, one could use the inverse function

$$\ln \left(\frac{y_{0.5}}{1 - y_{0.5}} \right) = \mu, \tag{7.7.16}$$

given any chosen median $y_{0.5} \in (0, 1)$ to compute μ. Then, by choosing a value $y_q \in (0, 1)$ for $q = 0.9750$, a value of σ could be computed by following a procedure similar to that above.

For any Gaussian mixing process $Z(s)$ on a set $s \in \mathbb{S}$ of n nucleotide sites with covariance function $\gamma(s_1, s_1)$, expectation $E[Z(s)] = \mu$ and variance $var[Z(s)] = \gamma(s, s) = \varsigma^2$, a logistic Gaussian process would be defined by using the transformation

$$Y(s) = \frac{\exp(Z(s))}{1 + \exp(Z(s))} \in (0, 1) \tag{7.7.17}$$

for $s = 1, 2, \ldots, n$. Whenever it is possible to exercise choices for the parameters of the Gaussian mixing process, the parameters could be chosen based on choices for the quantiles of the logistic Gaussian Y-process such as the median and other quantiles of interest illustrated in the above discussion. If one wishes to confine attention to rates in sub-intervals of $(0, 1)$ of the form (θ_1, θ_2), then linear transformations of the form

$$\Theta(s) = \theta_1 + (\theta_2 - \theta_1) Y(s) \tag{7.7.18}$$

for $s = 1, 2, \ldots, n$ could be used. If any quantile y_q for the Y-process is chosen, then it can be transformed to the interval (θ_1, θ_2) by using the transformation

$$\theta_q(s) = \theta_1 + (\theta_2 - \theta_1) y_q. \tag{7.7.19}$$

for $s = 1, 2, \ldots, n$.

Let $\mathbb{S}_k$ denote may subset of the set $\mathbb{S}$ of n sites for $2 \leq k \leq n$ and let

$$Y(\mathbb{S}_k) = (Y(s) \mid s \in \mathbb{S}_k) \tag{7.7.20}$$

a collection of random variables of the .logistic Gaussian process corresponding to the sites in the set $\mathbb{S}_k$. With a view toward deriving a formula for the joint density function of the collection $Y(\mathbb{S}_k)$ of random variables, let $\Gamma(\mathbb{S}_k)$ denote its covariance matrix. If this matrix is non-singular, then it would be possible to set down a formula of the joint density of the random variables in $Y(\mathbb{S}_k)$ by using a technique that is similar to that used above in deriving a formula for the density of the logistic normal distribution in the one dimensional case. However, the derivation of this formula will not be pursued here and will be left as an exercise for an interested reader.

In the next chapter, the ideas discussed in this chapter will be used to organize the presentation of a set of algorithms for computing realizations of a nucleotide substitution process for large number of sites on a DNA molecule.

7.8 Nucleotide Substitution in a Three Site Codon

Among the fundamental discoveries of molecular genetics was the realization that nucleotides at three consecutive sites, called codons, code for amino acids, the building blocks of proteins. A discussion of these discoveries may be found in many books on molecular evolution such as Li (1997) and Nei and Kumar (2000). Because each site may be occupied by any one of four bases, the number of possible codons is $4^3 = 2^6 = 64$.

There are, however, only 20 amino acids so that some of the possible codons code for the same amino acid. The standard or "universal" genetic codes are listed in Table 1.1 of Nei and Kumar and the one and 3-letter abbreviations of the 20 amino acids are listed in Table 1.2. It is also of interest to note that the codes for vertebrate mitochondrial DNA differ from those of the standard nuclear DNA and are listed in Table 1.3 of Nei and Kumar. Because messenger RNA is involved in the coding process, the symbol T for thiamine has been replaced by U, the symbol for uracil, in the codons listed in these tables.

Given this information on codons, it seemed appropriate to conduct a computer simulation experiment designed to study the effects of the mutational process of nucleotide substitution on the evolution of a three site codon. In particular, attention will be focused on the existence of homoplasy and back mutation in the simulated data. For this experiment the FOAR mixing process was used and the chosen parameters values were the same at those used in experiment I reported in the preceding section.

As attention will be focused only on the sequences of simulated mutations, the only computer output that will be displayed will be the four mutations that occurred in each of four replications of the process. The reason for choosing this small number of mutations and replications was the wish to display an illustrative sample of simulated data in a brief table. Throughout this experiment, the mitochondrial "Eve" was the three letter codon AUG, which codes for the amino acid Methionine with the three letter symbol Met. Presented in table 7.8.1 is a table of the simulated data.

Table 7.8.1 Four Replications of Four Nucleotide Substitutions

·	**Rep. 1**	**Rep. 2**	**Rep. 3**	**Rep. 4**
0	AUG	AUG	AUG	AUG
1	GUG	GUG	AUA	UUG
2	GUU	GUA	UUA	UUA
3	GGU	GUG	UUG	UUU
4	CGU	GCG	CUG	CUU

In the first column of this table, the symbol 0 represents the initial codon AUG and the symbols $1, 2, 3, 4$ correspond to the four mutations simulated in each of the four replications of the experiment. For example, in the column representing replication 1, the initial codon AUG was transformed to the codon GUG due to the nucleotide substitution $A \to G$. In this simulated data, each of the four replications may be thought of as populations evolving in different geographic locations. In this connection, it is interesting to note that in replications 1 and 2, the codon GUG appears as the first mutation, which would be an example of homoplasy. It is also of interest to note, in the column representing replication 2, that a back mutation occurred. By observing the column for replication 2 it can the seen that nucleotides substitutions $G \to A \to G$ occurred in mutations 2 and 3, resulting the codon transformations $GUG \to GUA \to GUG$. It is also of interest to note that among the four replications in the simulation experiment, only in replication 2 was there a case in which a mutant codon was transformed back to a codon that had appeared previously. There was, however, a case of back mutation at a single site in replication 3 of the experiment. For in this replication the nucleotide substitutions $G \to A \to G$ were observed in the third site of the codon.

This illustrative experiment on nucleotide substitutions is interesting, because it demonstrates the existence of the phenomena of homoplasy and back mutation even in a very small sample of simulated data, but this small sample was not large enough to draw any general conclusions regarding the prevalence of these phenomena in existing populations. To estimate the prevalence of these phenomena in a larger sample of simulated sequences of nucleotide substitutions in an experiment consisting of 22 mutations and 50 replications, it would be necessary to develop software such that the prevalence of homoplasy and back mutation among the 50 replications of 22 mutations could be counted with respect to 1,120 base sites under consideration. The designing of and the writing of such software presents formidable technical problems in writing code that will not be attempted here. But, nevertheless, it is of interest to briefly outline some ideas that may be helpful in writing such code.

In its present state of development, the sites among the 1,120 sites under consideration where each of the 22 mutations occur as well as the type of nucleotide substitution are saved in a file for each of the 50 replications. Therefore, within each replication, it would be possible to write software to determine whether there were some sites such that two or more mutations occurred, which would be candidates for the phenomenon of back mutation. Then, by searching the nucleotides in a sequence of mutations at a given site, it could be determined whether back mutations did indeed occur. By repeating this procedure for each of the 50 replications, it would be possible to use the simulated data to get an estimate of the prevalence of back mutations.

To get an estimate of the prevalence of homoplasy in a sample of simulated data, it would be necessary to write code such that it would be possible to use the data sets in which the sites where mutations did occur in each of the 50 replications. Then, by searching the intersections of these 50 sets the sites, it would, in principle, be possible to determine those sites at which mutations occurred in two or more replications. Those sites, if any, in these intersections, would be candidates for the phenomenon of homoplasy. Whether homoplasy did indeed occur among the 50 replications could be checked by examining sequences of nucleotide substitutions in those replications belonging to the intersections of sets of sites where mutations occurred. To simplify the writing, the ideas just outlined were based on 22 mutations with 50 replications for the sake of concreteness, but in the software the number of mutations and replications could be arbitrary integers ≥ 1.

7.9 Computer Simulation Experiments With a Logistic Gaussian Mixing Process

For experiments with the FOAR process reported in a previous section, it was assumed that the mutation rates for transitions and transversions were distributed uniformly in selected subintervals of $(0, 1)$. It is, therefore, natural to ask whether the results of nucleotide substitution process would change if the mutations rates followed some distribution other than the uniform. As was also demonstrated in a previous section, a logistic Gaussian mixing process provides a convenient framework for testing the effects of an experimenter's choice for the distributions governing transition and transversions rates on the distributions of waiting times among mutations as well as the distributions of the times required to accumulate some fixed number of mutations in a population. The procedure used in constructing a logistic Gaussian mixing process was that of simulating realizations of Gaussian FOAR process with a prescribed expectation μ and variance σ^2 and then mapping these realizations into the interval $(0, 1)$, using the logistic distribution function.

In order to simulate realizations of a Gaussian FOAR process with selected expectation μ and standard deviation σ, it is necessary to choose the parameters of the FOAR process such that each random variable has expectation 0 and standard deviation 1. As is well known, if σ_ϵ^2 is the variance of the Gaussian noise in a FOAR process, then the variance of any random variable in the process is

$$\frac{\sigma_\epsilon^2}{1-\beta^2}, \tag{7.9.1}$$

where β is the autoregressive parameter. Therefore, if $\sigma_\epsilon = \sqrt{1-\beta^2}$, then all random variables in the FOAR process have expectation 0 and standard deviation 1. Throughout all the three illustrative experiments reported in this section, the value assigned to autoregressive parameter was $\beta = 0.9$, and the number mutations simulated with respect to 1,120 sites of a DNA molecule in each of 50 replications was 22.

As was demonstrated in a previous section, the expectation μ and standard deviation σ for the FOAR process may be determined by choosing two quantiles of the logistic normal distribution. In experiment IV, the parameter μ was determined by selecting the median of the logistic normal distribution as $Q_{50} = 0.5$. Then, given this median, the parameter σ was determined by choosing quantile $Q_{97.5}$ as $Q_{97.5} = 0.75$. Similarly, in experiment V, Q_{50} was chosen as $Q_{50} = 0.25$ and $Q_{97.5}$ was chosen as $Q_{97.5} = 0.5$. In both these experiments, transition rates were assigned to the interval $\left[10^{-7}, 10^{-6}\right]$ and transversion rates were assigned to the interval $\left[10^{-8}, 10^{-7}\right]$.

In experiment VI, transition and transversion rates were chosen such that the simulation results would be more in agreement with the "Out of Africa" hypothesis. Thus, in this experiment transition rates were assigned to the interval $\left[10^{-7.1}, 10^{-7.1} + 2 \times 10^{-7}\right]$ and transversion rates were assigned to the interval $\left[10^{-7.6}, 10^{-7.6} + 2 \times 10^{-7.5}\right]$. To determine μ and σ for this experiment, the median of the logistic normal distribution was chosen at $q_{50} = 0.25$ and the 97.5 quantile was chosen as $q_{97.5} = 0.5$. Presented in the tables 7.9.1 and 7.9.2 are selected quantiles for the estimated distributions of waiting time among 22 mutations and the times taken to accumulate 22 mutations is a population for the three experiments under consideration.

Table 7.9.1 Quantiles of Pooled Waiting Times in Years Among 22 Mutations

Quan	**Exp *IV***	**Exp *V***	**Exp *VI***
Min	0.1793	0.3063	1.0881
Q_{25}	428.8170	705.0713	2,097.1528
Q_{50}	1,044.8606	1,755.0161	5,295.9773
Q_{75}	2,004.2640	3,333.6497	10,259.2557
Max	9,822.2261	16,279.7231	53,612.0184

From these two tables, it can be seen in experiments IV and V that the choice of values for 50 and 97.5 quantiles had a pronounced effects on not only the waiting times among mutations but also for the waiting times to accumulate 22 mutations in a population. For example, for experiment IV when $Q_{50} = 0.5$ for the logistic normal distribution, the Max statistics for the waiting times among mutations and the time to accumulate 22 mutations in a population were approximately 9,800 and 48,000 years, respectively. It is also of interest to note that these numbers are also close to

Table 7.9.2 Quantiles of Waiting Times in Years to Accumulate 22 Mutations in a Population

Quan	**Exp *IV***	**Exp *V***	**Exp *VI***
Min	17, 726.4612	29, 504.7797	95, 371.1603
Q_{25}	27, 899.0725	45, 534.1414	140, 082.8371
Q_{50}	32, 049.3541	53, 365.0813	158, 849.7129
Q_{75}	37, 064.5967	61, 630.5045	189, 535.7900
Max	48, 164.4271	79, 130.0368	244, 411.7564

those for experiment II reported in the section on applications of the FOAR process. However, when the median of the logistic normal distribution was lowered to $Q_{50} = 0.25$, the corresponding Max statistics in experiment V were about 16,000 and 79,000, respectively. Thus, as one might expect, when $Q_{50} = 0.25$ for the logistic normal distribution, rates for transitions and transversions would be skewed to the left of their ranges, resulting in larger times among mutations and times to accumulate mutations in a population.

In experiment VI, however, the effect of shifting the median to $Q_{50} = 0.25$ did not produce the same skewed effects as observed in experiments IV and V. For if one compares the statistics summarized in the two tables above with those for experiment III is the section on experiments with the FOAR mixing process, it can be seen that the waiting times among mutations and the times taken to accumulate 22 mutations in a population do not differ significantly. Moreover, in an experiment not reported here with $Q_{50} = 0.5$ and $Q_{97.5} = 0.75$ for the logistic normal distribution did not differ significantly from those displayed in the above tables for experiment VI. These results suggest that the intervals selected for rates of transitions and transversions that were in closer agreement with the "Out of Africa Hypothesis" seem to be quite robust with respect to choice of distribution an experimenter assigns to these intervals.

It appears that whether selected intervals for rates of mutations are robust with respect to distributions assigned to them may also depend on the relative lengths of these intervals. For example, the ratio of length of the intervals for transitions in experiment VI to that used in experiments IV and V is

$$\frac{2 \times 10^{-7}}{10^{-6} - 10^{-7}} = 0.2222, \tag{7.9.2}$$

when truncated to four decimal places, and the same ratio for transversions is

$$\frac{2 \times 10^{-7.5}}{10^{-7} - 10^{-8}} = 0.7027, \tag{7.9.3}$$

when truncated to four decimal places. Thus, the lengths of the intervals of rates used in experiment VI were only fractions of those used in experiments IV and V. It is also of interest to compare the lengths of the interval used for rates of transversions in experiments VI to that used for transitions in this experiment. This ratio is

$$\frac{2 \times 10^{-7.5}}{2 \times 10^{-7}} = 0.3162 \tag{7.9.4}$$

when truncated to four decimal places. Interestingly, the length of the interval for transversions is less than a third of that for transitions in experiment VI. These observations also suggest that the tightness of these intervals in experiment VI contributed the robustness with respect to choice of distribution of mutation rates for these intervals.

Bibliography

[1] Bain, L. J. and Englehardt, M. (1987) **Introduction to Probability and Mathematical Statistics**. Duxbury Press, Boston.

[2] Box, G. E. P. and Jenkins, G. M. (1976) **Time Series - Forecasting and Control**. Holden-Day, Oakland, California.

[3] Brillinger, D. R. (1981) **Time Series - Data Analysis and Theory**. Holden-Day, Inc. San Francisco.

[4] Brockwell, P. J. and Davis, R. A, (1991) **Time Series: Theory and Methods**. Spring-Verlag, New York, Berlin and London.

[5] Fuller, W. A. (1976) **Introduction To Statistical Time Series**. John Wiley and Sons, New York and London.

[6] Kennedy, W. J. and Gentle, J. E. (1980) **Statistical Computing**. Marcel Dekker, Inc., New York and Basel.

[7] Li, W. H. (1997) **Molecular Evolution**. Sinauer Associates Inc. Sunderland, Mass 01375.

[8] Loève, M. (1955) **Probability Theory**. D. Van Nostrand Company, Inc., Princeton. New Jersey, Toronto, New York and London.

[9] Lukacs, E. (1960) **Characteristic Functions**. Charles Griffin & Company, London.

[10] Muirhead, R. J. (1982) **Aspects of Multivariate Statistical Theory**. John Wiley & Sons, Inc., New York, Chichester, Brisbane, Toronto and Singapore.

[11] Nei, M. and Kumar, S. (2000) **Molecular Evolution and Phylogenetics**. Oxford University Press.

[12] Roussas, G. G. (1973) **A First Course in Mathematical Statistics**. Addison-Wesley Publishing Company, Reading, Mass, Menlo Park, London and Don Mills, Ontario.

[13] Thisted, R. A. (1988) **Elements of Statistical Computing-Numerical Computation**. Chapman Hall, New York and London.

Chapter 8

Computer Implementations and Applications of Nucleotide Substitution Models at Many Sites – Other Non-SNP Types of Mutation

8.1 Introduction

This chapter is a continuation of chapter 7 in which mixtures of Markov models with mixing processes derived from Gaussian processes were introduced to accommodate nucleotide substitutions at many sites. In keeping with a theme openness regarding the mathematics underlying any Monte Carlo simulation model, this chapter begins with an overview of algorithms used in the Monte Carlo implementations of the stochastic structure introduced in chapter 7. The objective of this overview is to present the technicalities in sufficient detail so that any interested investigator could, in principle, check the reported results of any Monte Carlo simulation experiment by writing code in a programming language of his choosing and repeating the experiment. In order to provide a setting for potential applications of the Monte Carlo simulation procedures described in section 2, an overview of genographic research project is provided in section 3. Briefly, this project consists of studies of human origins by classifying existing human populations into Haplogroups based on about 22 $SNP's$ that have been found in human mitochondrial DNA, $mtDNA$.

Most of these 22 $SNP's$ occur in a section of the human mitochondrial genome called the D-loop, which consists of about 1,120 base sites. Therefore, in order to provide a concrete example that would have potential applications in the genographic project, three Monte Carlo simulation experiments, in which the evolution of 22 mutations among 1,120 base sites were to be simulated, were designed and executed. Among the objectives of these experiments was to gain information as to whether replicated experiments of this type could be executed on a desk top computer within a reasonable lengths of time. A second objective was to gain insights into the ranges of rates in the interval $(0, 1)$ that would lead to plausible distributions for waiting times among mutations and the times taken to accumulate 22 mutations in the $mtDNA$ that would be consistent with the "Out of Africa" hypothesis of human origins currently held by many workers in the genographic project. Roughly, this hypothesis states that modern human populations arose from small waves of migrations out of Africa that occurred somewhere in the neighborhood of 200,000 years ago. Section 4 contains an account of three exploratory experiments, each of which took reasonable amount of time, about seven minutes, to complete on a desk top computer. In one of these experiments, rates were chosen in the interval $(0, 1)$ such that the estimated median time to accumulate 22 mutations in the D-loop was about 160,000 years, which was judged to be in plausible agreement with the "Out of Africa" hypotheses.

In the remaining sections of the chapter, various topics of interest for further research are discussed. In section 5, the results of an experiment with mutations in a codon, consisting of three sites, are reported in which back mutations and parallel mutations were observed among the four replications. Both these types of mutations complicate the problem of classifying populations into Haplogroups and some notes as to how to extend the software to estimate the prevalence of these two types of mutations are also included in this section. In order to test the robustness of the results reported in section 4 to the choice of rate distribution on intervals in $(0, 1)$, in section 6 some of the experiments reported in section 4 are repeated, using the condition that rates in intervals of $(0, 1)$ follow a logistic normal distribution rather than the uniform distribution used in section 4. Section 7 contains a discussion for a need to extend the software so that evolution of protein coding genes may be studied by Monte Carlo simulation methods. Finally, the chapter ends with some preliminary notes on constructing stochastic models of a class of mutations called indels as well as other types of non-SNP mutations which have been encountered in the sequencing of genomes in man and other species.

8.2 Overview of Monte Carlo Implementations for Nucleotide Substitution Models with N Sites

In chapter 7, two types of Gaussian mixing processes were introduced. The type discussed first was a class of processes that were defined by choosing a specific type of covariance function depending on two parameters. Basic to any Monte Carlo implementation of a model in this class was whether it would be practical to implement the Choleski factorization of a covariance matrix into a lower triangular matrix, given numerical specification of the two parameter covariance function and the number of sites $n \geq 1$ sites under consideration. A second class of processes involved time series models with the property that realizations of the process could be computed recursively, given specified parameter values. When implementing a Markov nucleotide substitution model with respect to $n \geq 1$ sites of a DNA molecule, it is required that three realizations of the Gaussian mixing process must be computed for each site, because for any given nucleotide i_s at any site s, the only substitutions that are possible are those to one of three nucleotides i_{ν_1}, i_{ν_2} and i_{ν_3} such that $i_{\nu_j} \neq i_s$ for $j = 1, 2, 3$. Consequently, if $n \geq 1$ sites are under consideration, then $3 \times n$ realizations of the Gaussian mixing process will be required to compute rates of substitution per unit time for the n sites. When $n \geq 1000$, the computation of the Choleski factorization of a $3n \times 3n$ covariance matrix would not be feasible for most present day desk top computers. Therefore, for models with a large number of sites, one needs to implement a time series model with the property that realizations of the Gaussian mixing process may be computed recursively.

For any choice of a Gaussian mixing process, let Z_j for $j = 1, 2, 3, \ldots, 3n$ denote $3n$ Monte Carlo realizations of the process, and let μ and $\sigma_Z = \sqrt{var\left(Z_j\right)}$ denote the common expectation and standard deviation. Then, to transform these realizations of the mixing process into realizations of uniform random variables on the interval $(0, 1)$, one would compute the random variables

$$U_j = \Phi\left(\frac{Z_j - \mu}{\sigma_Z}\right) \tag{8.2.1}$$

for $j = 1, 2, 3, \ldots, 3n$. In the software used in the implementation of these ideas in computer experiments that will be discussed subsequently, the standard normal distribution function $\Phi\left(\cdot\right)$ was approximated by a well known formula which is known to perform well. Suffice it to say, that this approximation was not the one attributed to Hastings mentioned in the book Thisted (1988). Given this collection of uniform random variables, the next step is to transform them into sets of three rates for each of the n sites under consideration.

To transform this set of uniform random variables into substitution rates at the n sites, it is necessary to specify the nucleotide at each site. Let

$$NS0 = \left(i_\nu \mid \nu = 1, 2, \ldots, n\right) \tag{8.2.2}$$

denote the initial set of nucleotides at the n sites. For every i_ν in the set $NS0$

$$i_\nu \in (A, G, C, T) \leftarrow (1, 2, 3, 4)\,, \tag{8.2.3}$$

where the symbols on the right indicate that the four bases in a set of sites are coded by the numbers $1, 2, 3, 4$ in the software. For purposes for testing the software, the set $NS0$ could be computed by choosing numbers at random from the set $(1, 2, 3, 4)$. Alternatively, $NS0$ could be an actual set of observed nucleotides in a set of n chosen sites of a DNA molecule. To initialize a simulation run, the array NS will be assigned $NS = NS0$.

The next step in the Monte Carlo implementation of a model with n sites is to transform the $3 \times n$ realizations of uniform random variables into nucleotide substitution rates for each of the n sites such that transitions and transversions will be accommodated. If, for example, for some site ν the nucleotide $i_\nu = A$, then the substitution $A \to G$ would be classified as a transition, because another purine G was substituted for the purine A. Whereas, for either of the substitutions $A \to C$ or $A \to T$, a pyrimidine would be substituted for a purine so that such substitutions would be classified as transversions. Similarly, substitutions by a pyrimidine for another pyrimidine are called transitions and substitutions of purines by a pyrimidine are called a transversions.

To take into account the idea that transition rates are higher than transversion rates, let $\delta_{1trans} < \delta_{2trans}$ be two numbers in $(0, 1)$ be the bounds for transition rates and let the two numbers $\delta_{1transv} < \delta_{2transv}$ be the bounds for transversion rates such that $\delta_{2transv} \leq \delta_{1trans}$. To illustrate these ideas for the example under consideration, for some site ν let $U_{\nu j}$ for $j = 1, 2, 3$ denote the three uniform random variables computed for site ν. Then, for the substitution $A \to G$, the transition rate would be computed as

$$\theta_{\nu AG} = \delta_{1trans} + (\delta_{2trans} - \delta_{1trans})\, U_{\nu 1}\,. \tag{8.2.4}$$

Similarly, for the substitution $A \to C$, the transversion rate would be computed as

$$\theta_{\nu AC} = \delta_{1transv} + (\delta_{2transv} - \delta_{1transv})\, U_{\nu 2}\,. \tag{8.2.5}$$

The transversion rate $\theta_{\nu AT}$ would also be computed using this formula for the third random variable $U_{\nu 3}$. Briefly, by way of an overview on the part of this part of the Monte Carlo implementation procedure, software has been written to compute transition and transversion rates for each of the

n sites such that at site ν these rates take into account the nucleotide i_ν at site ν.

The step that follows the computation of rates of nucleotide substitutions for the n sites is that of computing realizations of the sojourn or holding times for each site. For site ν, let $\theta_{\nu ij}$ for $j \neq i$ denote the three substitution rates. Then, the parameter in the exponential holding time distribution for this site is

$$\theta_\nu = \sum_{j \neq i} \theta_{\nu ij}. \tag{8.2.6}$$

Let U_ν for $\nu = 1, 2, \ldots, n$ be a collection of independent uniform random variables on the interval $(0, 1)$ which would be computed independent of those discussed above, and let T_ν denote the holding or waiting time to the first mutation for the Markov substitution process at site ν. Then, the waiting time to the first mutation at each of the n sites would be computed using the formula

$$T_\nu = -\frac{1}{\theta_\nu} \ln (U_\nu) \tag{8.2.7}$$

for $\nu = 1, 2, \ldots, n$.

When all the n are considered simultaneously, the waiting time T to the first mutation, nucleotide substitution, would be the minimum of the waiting times for each of the sites. In symbols,

$$T = \min (T_\nu \mid 1 \leq \nu \leq n). \tag{8.2.8}$$

Observe that $T > t$ if, and only if, $T_\nu > t$ for all $\nu = 1, 2, \ldots, n$. Therefore, because these random variables are independent by assumption and exponentially distributed, it follows that

$$\begin{aligned} P[T > t] &= P[T_1 > t, T_2 > t, \ldots, T_n > t] \\ &= \prod_{\nu=1}^{n} P[T_\nu > t] = \exp(-\theta t), \end{aligned} \tag{8.2.9}$$

where

$$\theta = \sum_{\nu=1}^{n} \theta_\nu. \tag{8.2.10}$$

From this result it can be seen that the random variable T has an exponential random distribution with parameter θ and expectation

$$E[T] = \frac{1}{\theta}. \tag{8.2.11}$$

It may also be concluded from this formula that, because θ is an increasing function of n, the expected value $E\left[T\right]$ will decrease as n increases. One could use the observation that the random variable T has an exponential distribution to compute Monte Carlo realizations of T. But, for the Markov substitution process under consideration for the n sites, it is necessary to identify the site, say ν, where the mutation or substitution occurs so that a transition to another nucleotide at these site can be simulated.

To identify the site where the substitution will occur and also compute a realization of the random variable T, let

$$D = \left(T_\nu \mid \nu = 1, 2, \ldots, n\right) \tag{8.2.12}$$

denote the array of waiting times to the first mutation for the n sites. A useful way to compute a realization of the random variable T is to rank the elements of the array D from the smallest to the largest. Let R denote the array of ranks, and let $R\left[1\right]$ denote the first element of R. Then, $R\left[1\right]$ is the site where a nucleotide substitution will occur and realization of T may be computed by using the formula

$$T = D\left[R\left[1\right]\right], \tag{8.2.13}$$

where $D\left[R\left[1\right]\right]$ is the element in the array D at site or position $R\left[1\right]$.

To compute a realization of the process governing a substitution to another nucleotide, it will be helpful to consider a special case which will point to a general solution. Suppose, for example, $R\left[1\right] = \nu$ and $i_\nu = 1$. Then, the possible substitutions are $1 \to 2, 1 \to 3$ and $1 \to 4$. Let $\theta_{\nu 1 j}$ for $j = 2, 3, 4$ denote the computed rates of substitution., and let

$$p_{\nu 1 j} = \frac{\theta_{\nu 1 j}}{\theta_\nu} \tag{8.2.14}$$

denote the conditional probability of the transition $1 \to j$ for $j = 2, 3, 4$. Now consider a multinomial distribution with the probability vector

$$\mathbf{p}_\nu = \left(p_{\nu 1 j} \mid j = 2, 3, 4\right) \tag{8.2.15}$$

and suppose a sample of size one is computed using a Monte Carlo simulation procedure. Then, the computer would return one of three indicator vectors $\boldsymbol{\varepsilon}_1 = \left(1, 0, 0\right), \boldsymbol{\varepsilon}_2 = \left(0, 1, 0\right)$ and $\boldsymbol{\varepsilon}_3 = \left(0, 0, 1\right)$ with probabilities $p_{\nu 1 j}$ for $j = 2, 3, 4$. To compute which nucleotide substitution is realized, let

$$\boldsymbol{s} = \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix} \tag{8.2.16}$$

denote a 3×1 column vector, indicating the set of possible nucleotide substitutions. Then, the realized substitution would be given by the random inner product $\boldsymbol{\varepsilon}_j \boldsymbol{s}$, If, for example, $j = 1$, then $\boldsymbol{\varepsilon}_j \boldsymbol{s} = 2$. Given this illustrative example, it is clear that software could be written that would accommodate a nucleotide substitution from any base $i_v = 1, 2, 3, 4$. In the general case, let M denote the nucleotide that was realized in the substitution process. Then, the nucleotide sequence NS for the n sites would be modified by the symbolic transformation

$$NS\,[R\,[1]] \leftarrow M, \tag{8.2.17}$$

indicating that at site $R\,[1]$ in the nucleotide sequence NS the original nucleotide at this site was replaced by M.

At this point in a simulation experiment, the site $R\,[1]$ at which a mutation occurred, the waiting time T to the occurrence of this mutation and the modified nucleotide sequence NS have been computed such that the original base at site $R\,[1]$ in this sequence has been replaced with another base M. Because a stochastic process is being simulated, it is advisable that these steps be repeated a number of times so that some insights into the variability of the process may be gained. Suppose, for example, that the procedure just described was repeated $N_1 \geq 1$ times, resulting in an array of triplets

$$\left(\left(R\,[1]_k\,, T_k, NS_k\right) \mid k = 1, 2, \ldots, N_1\right). \tag{8.2.18}$$

It should be mentioned for $k = 1$, the initial nucleotide sequence is $NS = NS0$, and for $k = 2$, a form of NS modified by a nucleotide substitution at some site will be used as new initial sequence of nucleotides and so this process continues until N_1 mutations as simulated. At each step in this simulation, a total of $3 \times n$ realizations of the Gaussian mixing process are computed.

Given this array, it may also be of interest to consider realizations of the random variable

$$T_{N_1} = \sum_{k=1}^{N_1} T_k, \tag{8.2.19}$$

which would be interpreted as the time taken for N_1 mutations to accumulate in a sample of DNA molecules. However, when a study of a stochastic process using Monte Carlo simulation procedure is undertaken, it is advisable that more than one realization of any variable or array be computed.

One would thus be led to design an experiment such that N_2 replications of N_1 mutations are considered. In such an experiment, simulated data consisting of double arrays of the form

$$((R\,[1]_{kl}\,, T_{kl}, NS_{kl}) \mid k = 1, 2, \ldots, N_1; l = 1, 2, \ldots, N_2) \tag{8.2.20}$$

would be computed. From such simulated data, one could, for example, consider random variables of the form

$$T_{N_1 l} = \sum_{k=1}^{N_1} T_{kl} \tag{8.2.21}$$

for $l = 1, 2, \ldots, N_2$ as a sample of size N_2 of the times for N_1 mutations in a DNA molecule to accumulate. Other uses of such simulated data will be discussed when the results of several simulations experiments are discussed in the following sections of this chapter. At the beginning of each replication for $l = 1, 2, \ldots, N_2$, the initial nucleotide sequence is chosen as $NS0$ so that each computer run is a replication of a simulation experiment designed to compute N_1 nucleotide substitutions. It is of interest to note that the sites at which these substitutions occur will vary among the N_2 replications of the experiment.

In the foregoing discussion, it was assumed that realizations of the Gaussian mixing process were mapped into uniform random variables on the interval $(0, 1)$. However, this step in the procedure could easily be replaced by software that maps realizations of the Gaussian mixing process directly into logistic normal random variables, taking values in the interval $(0, 1)$, by computing

$$Y_j = \frac{\exp(Z_j)}{1 + \exp(Z_j)} \tag{8.2.22}$$

for $j = 1, 2, \ldots, 3n$. In this connection, two caveats should be mentioned. If one decided to let the quantiles of the logistic-Gaussian process be determined by the expectation μ and standard deviation σ_Z of the Gaussian mixing process, then the quantiles of the logistic Gaussian process would be determined by the equation

$$y_q = \frac{\exp(\mu + \sigma_Z z_q)}{1 + \exp((\mu + \sigma_Z z_q))}, \tag{8.2.23}$$

where z_q is the *q-th* quantile of the standard normal distribution. On the other hand, it one wished to determine the quantiles of the logistic Gaussian process by fixing the median at some value y_1 in $(0, 1)$ and another quantile at y_2 in $(0, 1)$ to determine the standard deviation, then for the Gaussian mixing process adjustments would need to be made such that $\mu = 0$ and $\sigma_Z = 1$. Examples of this procedure will be discussed in a subsequent section where the logistic Gaussian process will be used to compute rates of nucleotide substitutions.

8.3 Overview of Genographic Research Project – Studies of Human Origins

During work on the Human Genome Project, which was completed in 2000, new types of technology for sequencing genomes were developed, which led investigators in genetics and other disciplines to not only consider how these technologies could be used for future research in medicine and other fields but also how they could be utilized to elucidate the genetic history and origins of present day human populations. In 2005, the Genographic Project was launched by the National Geographic Society with the goal of collecting mitochondrial DNA in females and DNA from the Y chromosome of males with a view toward studying the genetic history of present day human populations. These two types of DNA are well suited for such studies, because mitochondrial DNA is passed from mothers to daughters and the selected DNA on the Y chromosome is passed from fathers to sons so that in both cases historical genetic patterns, which evolved over many generations, would not be complicated by process of genetic recombination. A description of the Genographic Project may be found in the interesting and informative book, Wells (2006), written for the general reader. The more technical paper by Behar *et al.* (2007) may be consulted for a description of the public participation mitochondrial DNA, mtDNA, data base that has been constructed from the DNA of individuals who volunteered for Genographic Project and consented to let their DNA be part of scientific research studies.

Reported in the paper of Behar *et al.* (2007) is the analysis of 78,590 samples of mtDNA of which 41,552, 5046, and 16,971, respectively, were genotyped with respect to a panel of 10, 20,21 and 22 single nucleotide polymorphisms, SNP's. Given this assignments of genotypes, individuals were classified into Haplogroups, Hg, using methodologies for inferring phylogenies. Simply put, Haplogroups are sets of people who are thought to share a common ancestry. A complete listing of these Hg along with their frequencies in the sample may be found in Behar *et al.* (2007) and a genographic classification of them based on phylogenetic trees has been discussed in detail in an appendix on Haplogroups in the book by Wells (2006). An extensive account of the statistical methods used for inferring phylogenetic trees, from which Hg's are defined, has been given in the interesting book Felsenstein (2004). The mutational process thought to be governing the evolution of SNP's is that of nucleotide substitution.

Two processes referred to as homoplasy and back mutations are thought to have complicated and rendered more uncertain the procedures for inferring phylogenetic trees. Homoplasy is the name of a phenomenon that is observed when two distinct branches of a phylogenetic tree exhibit the same mutation, i.e., the same mutation, defined as the presence of a nucleotide at the same site of mtDNA tree. The phenomenon of back mutation is said to have been observed when a mutation at some site used to identify a Hg but may have back mutated to the nucleotide that is thought to be the ancestral form. When these two phenomena are operational, two given haplotypes may be identical in state, i.e., display the same set of nucleotides, but not identical by descent, which may bias the classification of Hg's. In the sections that follow, some computer experiments designed to shed light on the phenomena of homoplasy, back mutation and the evolution of SNP's at some designated set of sites of a DNA molecule will be described and analyzed within the framework of stochastic nucleotide substitution process at multiple sites described in the preceding sections of this chapter.

Before proceeding to a discussion and analysis of computer experiments, however, an interesting paper of Ingman (2001) published on the world wide web is worthy of mention. In this paper, it is pointed out that most of the SNP's used to classify Hg's are located in the D-loop, which is about 7 percent of the mitochondrial genome and is typically 1,120 base pairs long. The phenomena of back mutation and homoplasy are also mentioned in this paper but instead of using the word, homoplasy, as defined above, an equivalent term, parallel substitution, is used. This latter term seems to be more suitable for modern English usage. Also mentioned in this paper is the possibility of mutational hot spots in the D-loop. In this connection, some computer experiments to be discussed in the sections that follow will shed light as to whether the alleged hot spots are indeed sites with higher rates of mutation or are merely due to the high levels of stochasticity which arise inherently in the DNA copying process connected with the division of cells within individuals and the passing of DNA from parents to offspring from generation to generation. Lastly, it is also suggested that the reader search the world wide web for more information on human mtDNA by using this title in a search engine. Among the many web sites that may arise in this search is one designated as MITOMAP, where complete sequences of the human mitochondrial genome are published along with annotations as to what genes in this genome may be implicated in human diseases. In passing, it is also of interest to note that mitochondria are the power plants of cell in the sense these bodies supply the enzymes that lead to the oxidation of sugars and other carbohydrates which release energy needed by a cell.

8.4 Simulating Nucleotide Substitutions in Evolutionary Time

As mentioned in the previous section, a typical number of base pairs in the D-loop of human mtDNA is 1,120. To implement a nucleotide substitution process accommodating that many sites, it will be necessary to choose a mixing process such that Monte Carlo realizations of it can be computed recursively so as to avoid the need to do a Choleski factorization of a large covariance matrix. The simplest case of a Gaussian mixing process, whose realizations may be computed recursively, is the first order autoregressive (FOAR) process introduced in a previous section. As was discussed in a previous section, a FOAR-process depends on three parameters namely; μ, the expectation of the process, β, the autoregressive parameter, and σ, the standard deviation of the Gaussian ϵ-process, which is sometimes called noise. In what follows trial values of these and other parameters will be selected and tested in preliminary experiments with nucleotide substitution model under consideration.

To conduct a Monte Carlo simulation experiment using a FOAR-process, numerical values of these parameters must be assigned. In all experiments reported in this section, these parameters values were chosen as $\mu = 0, \beta = 0.9$ and $\sigma = 2$. The rationale for choosing $\beta = 0.9$ was the thought that it would be of interest to investigate the performance of a FOAR-process for which rates on nucleotide substitution would be quite highly correlated for adjacent sites. A reason for choosing the value $\sigma = 2$ was the notion that copying and transmitting of DNA from generation to generation is a process with a moderate level of stochasticity. The expectation $\mu = 0$ was chosen so that realizations of the FOAR-process could be interpreted as deviations from their expected value.

The mitochondrial "Eve", $NS0$, chosen for the experiments reported in this section, was simulated on a computer by choosing $n = 1,120$ numbers, the number of base pairs in a typical D-loop of human mtDNA, at random with replacement from the set, $(1, 2, 3, 4)$, representing the four bases or nucleotides. The number of mutations or nucleotide substitutions simulated from the mitochondrial "Eve" in each replication of the experiment was $N_1 = 22$, the highest number of SNP's used to genotype samples of mtDNA from individuals reported by Behar *et al.* (2007). Moreover, each experiment consisting of $N_1 = 22$ mutations was replicated $N_2 = 50$ times. In each replication, $NS0$ was used as the mitochondrial "Eve" so that every replication was a Monte Carlo simulation experiment to simulate 22 nucleotide substitutions at the $n = 1,120$ sites of a DNA molecule.

As was expected, the intervals chosen in $(0, 1)$ for rates of transitions and transversions were very sensitive regarding the distribution of evolutionary times taken to simulate $N_1 = 22$ nucleotide substitutions. In experiment I, the interval in $(0, 1)$ chosen for rates of transitions had the bounds $\delta_{1trans} = 10^{-5}$ and $\delta_{2trans} = 10^{-4}$, and the bounds for rates of transversions were chosen as $\delta_{1transv} = 10^{-6}$ and $\delta_{2transv} = 10^{-5}$. To provide a basis for assessing the plausibility of these bounds, in experiment II transitions rates with bounds $\delta_{1trans} = 10^{-7}$ and $\delta_{2trans} = 10^{-6}$ were chosen and bounds for rates of transversions were chosen as $\delta_{1transv} = 10^{-8}$ and $\delta_{2transv} = 10^{-7}$. As will be illustrated in the preliminary experiments reported below, this choice of bounds led to a rather wide range of times taken to accumulate $N_1 = 22$ nucleotide substitutions in the D-loop of human mtDNA.

By way of an overview of the number of realizations of the FOAR-process computed in these experiments, let Z_j for $j = 1, 2, \ldots, 3 \times n$ denote the $3 \times n = 3,360$ Monte Carlo realizations of the FOAR-process that were used as the mixing process for each of the 22 mutations that were simulated in each of the 50 replications. Altogether, the total number of realizations of the FOAR-process computed in these experiments was

$$50 \times 22 \times 3360 = 3,696,000. \tag{8.4.1}$$

At each step in the experiment,

$$50 \times 22 \times 1120 = 1,232,\ 000 \tag{8.4.2}$$

calculations were made to generate realizations exponential random variables representing the sojourn times at each site. In the desk top computer used to do these calculations, the time taken to complete these over four million calculations was about seven minutes.

As was illustrated in previous sections, these realizations were transformed into correlated uniform random variables in the interval $(0, 1)$ by passing standized versions of them through a standard normal distribution function, and these uniform random numbers in turn were transformed into rates of transitions and transversions using the bounds given above.

A useful way of statistically summarizing a sample of Monte Carlo simulation data is to pool all the waiting times among mutations and compute the extreme values of this sample as well as a set of chosen quantiles. For example, the waiting time among the occurrence of 22 nucleotide substitutions were simulated for 50 replications which produced a pooled sample of size $22 \times 50 = 1,100$. The order statistics for this sample of $1,100$ observations were then computed, which made it possible to identify the minimum and maximum values in the data. Furthermore, the order statistics could also be used to compute the quantiles Q_{25}, Q_{50} and Q_{75} of the simulated data.

Presented in the tables 8.4.1 and 8.4.2 are these statistical summaries of experiments I and II for the pooled waiting times in years among the $1,100$ mutations simulated in each of experiments I and II.

Table 8.4.1 Quantiles of Pooled Waiting Times in Years Among 22 Mutations

Quan	**I**	**II**
Min	0.0016	0.1647
Q_{25}	4.5608	456.0800
Q_{50}	9.9005	990.0583
Q_{75}	20.0223	$2,002.2322$
Max	103.9689	$10,331.3111$

Another perspective from which the simulated Monte Carlo data may be viewed is to look at statistical summaries of the waiting times to the accumulation of 22 mutations in a DNA segment of $1,120$ base sites. These cumulative times were computed for each of the 50 replications in experiments I and II and the statistical summaries of this sample of size 50 are presented in the table below.

Table 8.4.2 Quantiles of Waiting Times in Years to Accumulate 22 Mutations

Quan	**I**	**II**
Min	161.4156	$16,141.5611$
Q_{25}	275.6035	$27,560.3547$
Q_{50}	309.1349	$30,913.4981$
Q_{75}	377.1201	$37,712.0109$
Max	499.9689	$49,996.8906$

A current "Out of Africa Hypothesis" on the origins of modern humans states roughly that modern humans originated in Africa about 200,000 years ago and since that time have spread around the world. Therefore, according to this hypothesis, mutations in the mtDNA used to classify present day humans into Haplogroups of related individuals evolved during a period of about 200,000 years. In this connection, the statistical summaries of the simulated data presented in Tables 8.4.1 and 8.4.2 suggest that the chosen rates of transitions and transversions were too high in experiments I and II, because even the Max estimates for times among mutations and the times to accumulate 22 mutations are too low.

For example, from Table 8.4.2 it can be seen that for experiment I the Max value for the time to accumulate 22 mutations was 499.9689 years and the value of this statistic for experiment II was $49,996.8906$ years, which are far short of 200,000 years. These observations in turn suggest that it would be of interest to do a simulation experiment in which the times to accumulate 22 mutation were in closer agreement with the "Out of Africa Hypothesis".

To get a handle on a range of mutation rates that could be chosen to satisfy the condition that about 200,000 years would be a sufficient time for a population to accumulate 22 mutations, it will be helpful to consider the expected time to accumulate this number of mutations for Markov nucleotide substitution process under consideration. Let 10^{-z} denote a constant trial mutation rate for each of n sites under consideration, where $z > 0$. Then, as explained in a previous section, it follows that the time T to the first mutation follows an exponential distribution with parameter

$$q = \sum_{\nu=1}^{n} 10^{-z} = n10^{-z} \tag{8.4.3}$$

and expectation

$$E\left[T\right] = \frac{1}{n10^{-z}} = \frac{10^{z}}{n}. \tag{8.4.4}$$

This formula can also be interpreted as the expected waiting times among the 22 mutations under consideration.

Let the random variables T_k for $k = 1, 2, \ldots, 22$ denote the waiting times among the 22 mutations. Then, the waiting time for the accumulation 22 mutations in the DNA is given by the random variable

$$T_{22} = \sum_{k=1}^{22} T_k, \tag{8.4.5}$$

where $E\left[T_k\right] = 10^z/n$ for all $k = 1, 2, \ldots, 22$. Therefore,

$$E\left[T_{22}\right] = \sum_{k=1}^{22} E[T_k] = \frac{22 \times 10^z}{n}. \tag{8.4.6}$$

For $n = 1,120$ and $z = 7$, a call to the computation engine yields the result

$$\frac{22 \times 10^7}{1120} = 196,428.\,571\,42 \tag{8.4.7}$$

years, when truncated to five decimal places. This calculation suggests that, because $E\left[T_{22}\right]$ is about $200,000$ years, a plausible rate of mutation for each of he $1,120$ sites under consideration is 10^{-7}.

A useful approach for selecting an interval around 10^{-7} for rates of transitions is to suppose 10^{-7} is the midpoint of this interval. Thus, the bounds for rates of transitions, should be chosen such that

$$\frac{\delta_{2trans} - \delta_{1trans}}{2} = 10^{-7}, \tag{8.4.8}$$

which implies that

$$\delta_{2trans} = \delta_{1trans} + 2 \times 10^{-7}. \tag{8.4.9}$$

Hence, if $\delta_{1trans} = 10^{-7.1}$, then

$$\delta_{2trans} = 10^{-7.1} + 2 \times 10^{-7}. \tag{8.4.10}$$

Similarly, if $10^{-7.5}$ is chosen as the midpoint of the interval for transversions and $\delta_{1transv} = 10^{-7.6}$, then

$$\delta_{2transv} = 10^{-7.6} + 2 \times 10^{-7.5}. \tag{8.4.11}$$

In computer experiment III, these intervals were used for transitions and transversions. Presented in the table below are the selected quantiles and extreme values for waiting times among mutations and the times taken to accumulate 22 mutations.

The summary statistics presented in Table 8.4.3 are in better agreement with the "Out of Africa Hypothesis" than those presented in Tables 8.4.1 and 8.4.2. Interestingly, although the formulation of the nucleotide substitution process under consideration does not take into account the population dynamics involving the effects of chance in the passing of DNA from parents to offspring from generation to generation, the Min value in Table 8.4.3 for waiting times among mutations is about a year, which may be sufficiently short time for a female carrier in which this mutation occurs to pass it on to her daughters with positive probability.

Table 8.4.3 Quantiles for Waiting Times in Years for Experiment III

Quan	Times Among Mutations	Times To Accumulate 22 Mutations
Min	1.0881	95,371.1727
Q_{25}	2,097.1527	140,082.8527
Q_{50}	5,296.9781	158,849.7264
Q_{75}	10,259.2564	189,535.8042
Max	53,613.0281	244,411.7711

With regard to the times to accumulate 22 mutations in the DNA, it is interesting to note that Min statistic is a little less than $10,000$ years, which suggests in some populations the time to accumulate 22 mutations can be relatively short in terms of evolutionary time. As can be seen from this table, the median quantile, Q_{50}, of the times to accumulate 22 mutation was nearly $159,000$ years; while the Max statistic was about $244,000$ years. Given the estimates in Table 8.4.3, it seems plausible that the ranges for rates of transitions and transversions are in pretty good agreement with the "Out of Africa Hypothesis" of about 200,000 years.

The Min statistics for the waiting times among mutations indicate that this value was observed for only one of the simulated times in the three experiments under consideration. But it is also of interest to get an estimate of the number of times among mutations that were less than or equal to about 20 years. For if a mutation occurs in the mitochondria of the cells destined to develop into eggs within a span to 20 years following her birth, a female may pass on this mutation to her female offspring with a positive probability and these offspring may in turn pass the mutation to their daughters.

Thus, after many generations, the mutation may become prevalent in a subpopulation that are descendants of the female in which the original mutation occurred. To provide some insights into how many times among the pooled 1,100 observations were $\leq$ 20 years, a software routine was written to count the number of simulated observations that were $\leq$ 20 years. For experiments I, II and III, these counts were $824, 19$ and 3, respectively.

It is interesting to note that in experiment III, which produced results in fairly good agreement with the "Out of Africa Hypothesis", the estimated probability that a mutation would occur in sufficient time to enter a germ line following the birth of a female was $3/1100 = 2.727\,272\,727\,272\,73 \times 10^{-3}$. In a subsequent chapter in which the stochastic population dynamics of the survival of mutations are discussed, the implications of such estimate will be further analyzed.

Table 8.4.4 Estimated Distribution of the Number of Mutations per Site for Experiments II and III

Sites with x Mutations	**Exp. II**	**Exp. III**
0	434	420
1	388	407
2	206	209
3	72	64
4	17	18
5	2	1
6	1	1

Another view of the simulated data may be obtained by counting the number of mutations that occurred at each of the 1,120 sites. It was possible to program a computer to do these counts, because the software has been designed to record the site of each mutation. Presented in table 8.44 is a table of these counts for the simulated data in experiments II and III. In column 1 of this table x denotes the number of sites among the 1,120 sites under consideration at which $x = 0, 1, 2, \ldots, 6$ mutations were counted. The numbers of sites with x mutations are listed in columns 2 and 3 for experiments II and III. For example in experiments II and III there were 434 and 420 sites, respectively, at which no mutations occurred, and there were 388 and 407 sites, respectively, in which only one mutation occurred. For all those sites for which there were $x \geq 2$ mutations, the phenomenon of back mutation would have been a possibility. From table 8.4.4 it can be seen that the fraction of sites such that $x \geq 2$ was $298/1120 = 0.266,\ 071$ and that for experiments III was $293/1120\ = 0.261,\ 607$. These numbers suggest that if the simulated data were analyzed in more detail, it seems likely that some back and parallel mutations may have occurred in these experiments. In the next section, the issue of back mutations will be examined in experiments in which only three sites were considered.

This simulated data in experiment III also suggests that the idea that some regions of a DNA molecule are hot spots for mutation should be approached with caution, because of the level of high stochasticity that may exist at the molecular level. In addition to the simulation experiment involving three letter codons, an overview of the software which was developed to count back and parallel mutations in the simulated data for experiment III will also be given in the next section.

8.5 Counting Back and Parallel Mutations in Simulated Data

Among the fundamental discoveries of molecular genetics was the realization that nucleotides at three consecutive sites, called codons, code for amino acids, the building blocks of proteins. A discussion of these discoveries may be found in many books on molecular evolution such as Li (1997) and Nei and Kumar (2000). As each site may be occupied by any one of four bases, the number of possible codons is $4^3 = 2^6 = 64$. There are, however, only 20 amino acids so that some of the possible codons code for the same amino acid. The standard or "universal" genetic codes are listed in their Table 1.1 of Nei and Kumar and the one and 3-letter abbreviations of the 20 amino acids are listed in their Table 1.2. It is also of interest to note that the codes for vertebrate mitochondrial DNA differ from those of the standard nuclear DNA and are listed in Table 1.3 of Nei and Kumar. Since messenger RNA is involved in the coding process, the symbol T for thiamine has been replaced by U, the symbol for uracil, in the codons listed in these tables.

Given this information on codons, it seemed appropriate to conduct a computer simulation experiment designed to study the effects of the mutational process of nucleotide substitution on the evolution of a three site codon. In particular, attention will be focused on the existence of homoplasy, parallel, and back mutation in the simulated data. For this experiment the FOAR mixing process was used and the chosen parameters values were the same at those used in experiment I reported in the preceding section. Because attention will be focused only on the sequences of simulated mutations, the only computer output that will be displayed will be the four mutations that occurred in each of four replications of the process. The reason for choosing this small number of mutations and replications was the wish to display an illustrative sample of simulated data in a brief table. Throughout this experiment, the mitochondrial "Eve" was the three letter codon AUG, which codes for the amino acid Methionine with the three letter symbol Met. Presented in Table 8.5.1 is a table of the simulated data.

In the first column of this table, the symbol 0 represents the initial codon AUG and the symbols $1, 2, 3, 4$ correspond to the four mutations simulated in each of the four replications of the experiment. For example, in the column representing replication 1, the initial codon AUG was transformed to the codon GUG due to the nucleotide substitution $A \to G$. In

Table 8.5.1 Four Replications of Four Nucleotide Substitutions

·	**Rep. 1**	**Rep. 2**	**Rep. 3**	**Rep. 4**
0	AUG	AUG	AUG	AUG
1	GUG	GUG	AUA	UUG
2	GUU	GUA	UUA	UUA
3	GGU	GUG	UUG	UUU
4	CGU	GCG	CUG	CUU

this simulated data, each of the four replications may be thought of as populations evolving in different geographic locations. In this connection, it is interesting to note that in replications 1 and 2, the codon GUG appears as the first mutation, which would be an example of homoplasy. It is also of interest to note, in the column representing replication 2, that a back mutation occurred. By observing the column for replication 2 it can the seen that nucleotides substitutions $G \to A \to G$ occurred in mutations 2 and 3, resulting the codon transformations $GUG \to GUA \to GUG$. It is also of interest to note that among the four replications in the simulation experiment, only in replication 2 was there a case in which a mutant codon was transformed back to a codon that had appeared previously. There was, however, a case of back mutation at a single site in replication 3 of the experiment. For in this replication the nucleotide substitutions $G \to A \to G$ were observed in the third site of the codon.

This illustrative experiment on nucleotide substitutions is interesting, because it demonstrates the existence of the phenomena of homoplasy and back mutation even in a very small sample of simulated data, but this small sample was not large enough to draw any general conclusions regarding the prevalence of these phenomena in existing populations. To estimate the prevalence of these phenomena in a larger sample of simulated sequences of nucleotide substitutions in an experiment consisting of 22 mutations and 50 replications, it would be necessary to develop software such that the frequencies of back and parallel mutations among the 50 replications of 22 mutations could be estimated with respect to 1,120 base sites under consideration. Fortunately, by a process of trial and error involving searches of the simulated data, it was possible to write such code, whose structure is outlined below. The definitions that follow are with respect to the data simulated in experiment III. Moreover, to cast this experiment in an evolutionary context, the 50 replications of the experiment will be viewed as 50 different lines of descent evolving from the same mitochondrial Eve $NS0$ with 1,120 simulated nucleotides.

Let $i \in NS0$ denote the initial nucleotide at some site and suppose there are two or more mutations at that site in some replication. For the sake of concreteness, suppose there are two mutations in some replication. Let the nucleotides i_1 and i_2 denote these mutations and suppose the nucleotide substitutions that led to them were the substitution $i \to i_1$ followed by substitution $i_1 \to i_2$. If $i_2 = i$, then i_2 would be classified as a back mutation. Observe that back mutations can be observed only at those sites in which there were two or more mutations in one replication, representing a particular line of descent. It could also happen that there may be more than two mutations in some replication. To illustrate this idea, let i denote the initial nucleotide at some site and let i_1, i_2 and i_3 denote the mutated nucleotides in three replications. Then, if either $i_2 = i$ or $i_3 = i$, then either i_2 or i_3 would be classified as a back mutation. Let S_b denote the set of all back mutations among the set S of $50 \times 22 = 1100$ mutations that were simulated.

On the other hand, parallel mutations refer to mutations in different replications, representing different lines of descent. For example, let $i \in NS0$ denote the initial nucleotide at some site and suppose at this site there are two or more mutations which occur in different replications. Parallel mutations are defined only with respect to an initial nucleotide i at some site. For the sake of illustration, suppose there are two mutations at some site and suppose they occur in different replications. Let $i \to i_1$ be the mutation in one replication and let $i \to i_2$ denote the mutation in the other replication. If $i_1 = i_2$ the two mutations in different replications are called parallel mutations. In general, if a site has some number $x \geq 2$ mutations in different replications, then a mutation in some replication belongs to the set of parallel mutations at this site, if this particular mutation occurs in two or more replications. Let S_p denote the set of parallel mutations among the set of 1100 mutations that were simulated.

The simulated data $\boldsymbol{D}$ were stored in the computer in the form of a three dimensional array with dimensions $50 \times 22 \times 1120$, representing 50 replications of 22 simulated mutations at 1,120 sites of a DNA molecule. The objective of the software discussed below was to provide a means for finding and counting the number of mutations that belonged to each of the sets S_b and S_p. The first step in developing this software was to write code that would partition the set of $1,120$ sites into disjoint sets $S_0, S_1, S_2 \ldots$ such that S_ν was the set of sites that contained ν mutations. As can be seen from Table 8.4.4, the number ν ranged over the values $\nu = 0, 1, 2, \ldots, 6$

in the simulated data of experiment III. Furthermore, because the focus of attention was to find back and parallel mutations, it sufficed to limit the search to those sets of sites with 2 or more mutations; namely the set of sites in the union

$$S_{2+} = \bigcup_{v=2}^{6} S_{\nu}. \tag{8.5.1}$$

The next step in the development of the software was to write code to screen the replications for each site $s \in S_{2+}$ to find one or more replications, among 50 replications that were simulated, where the mutations occurred. This screening procedure was based on counting the number of times the initial nucleotide i at some site s occurred among the 22 nucleotide substitutions that were simulated in each replication. For example, let $\boldsymbol{j} = (j_1, j_2, \ldots, j_{22})$ denote the set of 22 simulated nucleotides in some replication and compute the Boolean vector

$$\boldsymbol{\xi} = (i = \boldsymbol{j}) = (\xi_1, \xi_2, \ldots, \xi_{22}), \tag{8.5.2}$$

where $\xi_\nu = 1$ if $i = j_\nu$ and $\xi_\nu = 0$ if $i \neq j_\nu$ for $\nu = 1, 2, \ldots, 22$. Then, compute the sum

$$\delta = \sum_{\nu=1}^{22} \xi_\nu. \tag{8.5.3}$$

If $\delta = 22$, then there were no mutated nucleotides in the replication, but if $\delta < 22$, then there was at least one mutated nucleotide in the replication. Given this information, it was possible to find the replication or replications among the 50 replications where the mutations occurred.

The next step in the development of the software was to count the numbers of back and parallel mutations in the simulated data. This counting process will be illustrated for a site where it was known that there were 2 mutations. Let $s \in S_2$ and for this site let $R(s)$ denote the set of replications where the 2 mutations occurred, and let $\boldsymbol{D}[R(s); 22; s]$ denote the sub-array of $\boldsymbol{D}$ that contains the 2 mutations. If the dimensions of $\boldsymbol{D}[R(s); 22; s]$ are 2 by 22, then the two mutations would occur in two different replications and would thus be candidates for a test to determine whether they were or were not parallel mutations. If this array has dimensions 1 by 22 however, this is a signal that both mutations occurred in the same replication and would thus be eligible to a test for a back mutation. By using this criterion of array dimensions, it was possible to partition the set S_2 of sites into two sub-sets S_{21} and S_{22}, indicating whether the two mutations occurred in one or two replications. By summing over sites $s \in S_{21}$, it was possible to count the number of back mutations in the set S_{21}, and, similarly, by summing over the sites $s \in S_{22}$, it was possible to count the number of parallel mutations in the set S_{22}.

By using the dimension of array criterion and the simulated data for experiment III, it was also possible to partition the set of sites S_3, where there were three mutations, into two sub-sets. The sub-set S_{32} denotes the sub-set of sites in S_3 such that three mutations occurred in two replications, and the sub-set S_{33} denotes the set of sites in S_3 such that the three mutation occurred in three replications. For all sites in the set S_{32}, it was necessary to search and count both replications for back mutations as well as parallel mutations, and for sites in the set S_{33}, it sufficed to search for and count only parallel mutations. As it turned out, for all the sites in the set S_4, the four mutations turned out to be in four different replications so that it sufficed to find and count only parallel mutations. From table 8.4.4, it can be seen that each of the sets of sites S_5 and S_6 contained only one site. For the case of S_5, the five mutations occurred in five replications so that it sufficed to find and count only parallel mutations. Interestingly, the six mutations for the site in S_6 occurred in five replications so it was necessary to search for a back mutation in the replication that contained two mutations. Following this operation, all five replications were searched for parallel mutations, which were also counted. Presented in Table 8.5.2 are the counts of the numbers of back and parallel mutations in the simulated data of experiment III by using the software outlined above.

Table 8.5.2 Counts of Back and Parallel Mutations in Simulated Data of Experiment III

Back Mutations	8
Parallel Mutations	333

As can be seen from this table, among the 1,100 nucleotide substitutions simulated, there were 8 back mutations with a frequency of $8/1100 = 0.007273$ and 333 parallel mutations with a frequency of $333/1100 = 0.302\,7$. These results suggest that when developing a classification of Haplogroups based on $SNP's$ located in the D-loop of human mitochondrial DNA with 1,120 base pairs, it would be helpful to take into account the risk of errors in classification due to the presence of back and parallel mutations in a sample of lines of descent under consideration. Given these estimates of the frequencies of back and parallel mutation, one may conclude that back mutations present a significant smaller risk of misclassification than that for parallel mutations.

By way of an illustrative example, suppose there was no merging of lines of descent during the period of time taken to accumulate 22 mutations in each of the 50 lines, replications, under consideration. Moreover, suppose that using the "Out of Africa" hypothesis, one wanted to test whether some existing group on some continent had ancestors who passed through some point on another continent. A specific example of such an existing population is that of the aborigines of Australia for which it is thought that its ancestors passed through the southern tip of India on their way to Australia during the last 60 thousand years. Given this hypothesis, it seems plausible that some of the ancestors of the Australian aborigines may have stayed in southern India.

Now suppose the Australian aborigines carry a certain marker and this marker is also found in some present day Indian population and it is known that this population's ancestors have been in India for a very long period of time. Of course, it is tempting to conclude that the population in India and that of the Australian aborigines have common ancestors. On the other hand, the population in India may have descended from a different line carrying a parallel mutation who immigrated into India after the ancestors of Australian aborigines. According to the calculations outlined above, the probability of this misclassification is about 0.3 and the probability of a correct classification is about 0.7, under the assumption that there was no merging of the 50 simulated lines of descent. In view of this illustrative example, it is suggested that the methods of simulating lines of descent descending from a mitochondrial Eve $NS0$ will be useful in assessing probabilities of misclassifications of Haplogroups due to back and parallel mutations.

There are, of course, other types of uncertainties that arise when one considers classifying Haplogroups based on markers in the mitochondrial DNA. One source of uncertainty is that as lines of descent spread across the globe, during the migration out of Africa, there could have been a mixing and merging of lines. There are also issues concerning the concentration of markers in the D-loop of mitochondrial DNA, which contain only about 1,120 base sites. With this relatively small number of base sites, it is not surprising that when 1,100 mutations are simulated in an experiment, both back and parallel mutations were found in the simulated data. On the other hand, if the number of base sites were twice the number considered,

$1120 \times 2 = 2240$, in a simulation experiment, one would expect to find lower frequencies of both back and parallel mutations. It would be of interest to conduct such simulation experiments, but no further work in this direction will be pursued here.

8.6 Computer Simulation Experiments With a Logistic Gaussian Mixing Process

For experiments with the FOAR process reported in a previous section, it was assumed that the mutation rates for transitions and transversions were distributed uniformly in selected subintervals of $(0,1)$. It is, therefore, natural to ask whether the results of nucleotide substitution process would change if the mutations rates followed some distribution other than the uniform. As was also demonstrated in a previous section, a logistic Gaussian mixing process provides a convenient framework for testing the effects of an experimenter's choice for the distributions governing transition and transversions rates on the distributions of waiting times among mutations as well as the distributions of the times required to accumulate some fixed number of mutations in a population. The procedure used in constructing a logistic Gaussian mixing process was that of simulating realizations of Gaussian FOAR process with a prescribed expectation μ and variance σ^2 and then mapping these realizations into the interval $(0,1)$, using the logistic distribution function.

In order to simulate realizations of a Gaussian FOAR process with selected expectation μ and standard deviation σ, it is necessary to choose the parameters of the FOAR process such that each random variable has expectation 0 and standard deviation 1. As is well known, if σ_ϵ^2 is the variance of the Gaussian noise in a FOAR process, then the variance of any random variable in the process is

$$\frac{\sigma_\epsilon^2}{1-\beta^2}, \tag{8.6.1}$$

where β is the autoregressive parameter. Therefore, if $\sigma_\epsilon = \sqrt{1-\beta^2}$, then all random variables in the FOAR process have expectation 0 and standard deviation 1. Throughout all the three illustrative experiments reported in this section, the value assigned to autoregressive parameter was $\beta = 0.9$, and the number of mutations simulated with respect to 1,120 sites of a DNA molecule in each of 50 replications was 22.

As was demonstrated in a previous section, the expectation μ and standard deviation σ for the FOAR process may be determined by choosing two quantiles of the logistic normal distribution. In experiment IV, the parameter μ was determined by selecting the median of the logistic normal distribution as $Q_{50} = 0.5$. Then, given this median, the parameter σ was determined by choosing quantile $Q_{97.5}$ as $Q_{97.5} = 0.75$. Similarly, in experiment V, Q_{50} was chosen as $Q_{50} = 0.25$ and $Q_{97.5}$ was chosen as $Q_{97.5} = 0.5$. In both these experiments, transition rates were assigned to the interval $\left[10^{-7}, 10^{-6}\right]$ and transversion rates were assigned to the interval $\left[10^{-8}, 10^{-7}\right]$.

In experiment VI, transition and transversion rates were chosen such that the simulation results would be more in agreement with the "Out of Africa" hypothesis. Thus, in this experiment transition rates were assigned to the interval $\left[10^{-7.1}, 10^{-7.1} + 2 \times 10^{-7}\right]$ and transversion rates were assigned to the interval $\left[10^{-7.6}, 10^{-7.6} + 2 \times 10^{-7.5}\right]$. To determine μ and σ for this experiment, the median of the logistic normal distribution was chosen at $Q_{50} = 0.25$ and the 97.5 quantile was chosen as $Q_{97.5} = 0.5$ Presented in the tables below are selected quantiles for the estimated distributions of waiting time among 22 mutations and the times taken to accumulate 22 mutations in a population for the three experiments under consideration.

Table 8.6.1 Quantiles of Pooled Waiting Times in Years Among 22 Mutations

Quan	Exp IV	Exp V	Exp VI
Min	0.1793	0.3063	1.0881
Q_{25}	428.8170	705.0713	2,097.1528
Q_{50}	1,044.8606	1,755.0161	5,295.9773
Q_{75}	2,004.2640	3,333.6497	10,259.2557
Max	9,822.2261	16,279.7231	53,612.0184

Table 8.6.2 Quantiles of Waiting Times in Years to Accumulate 22 Mutations in a Population

Quan	Exp IV	Exp V	Exp VI
Min	$17,726.4612$	$29,504.7797$	$95,371.1603$
Q_{25}	$27,899.0725$	$45,534.1414$	$140,082.8371$
Q_{50}	$32,049.3541$	$53,365.0813$	$158,849.7129$
Q_{75}	$37,064.5967$	$61,630.5045$	$189,535.7900$
Max	$48,164.4271$	$79,130.0368$	$244,411.7564$

From these two tables 8.6.1 and 8.6.2, it can be seen in experiments IV and V that the choice of values for 50 and 97.5 quantiles had pronounced effects on not only the waiting times among mutations but also for the waiting times to accumulate 22 mutations in a population. For example, for experiment IV when $Q_{50} = 0.5$ for the logistic normal distribution, the Max statistics for the waiting times among mutations and the time to accumulate 22 mutations in a population were approximately 9,800 and 48,000 years, respectively. It is also of interest to note that these numbers are also close to those for experiment II reported in the section on applications of the FOAR process. However, when the median of the logistic normal distribution was lowered to $Q_{50} = 0.25$, the corresponding Max statistics in experiment V were about 16,000 and 79,000, respectively. Thus, as one might expect, when $Q_{50} = 0.25$ for the logistic normal distribution, rates for transitions and transversions would be skewed to the left of their ranges, resulting in larger times among mutations and times to accumulate mutations in a population.

In experiment VI, however, the effect of shifting the median to $Q_{50} = 0.25$ did not produce the same skewed effects as observed in experiments IV and V. For if one compares the statistics summarized in the two tables above with those for experiment III in the section on experiments with the FOAR mixing process, it can be seen that the waiting times among mutations and the times taken to accumulate 22 mutations in a population do not differ significantly. Moreover, in an experiment not reported here with $Q_{50} = 0.5$ and $Q_{97.5} = 0.75$ for the logistic normal distribution did not differ significantly from those displayed in the above tables for experiment VI. These results suggest that the intervals selected for rates of transitions and transversions that were in closer agreement with the "Out of Africa Hypothesis" seem to be quite robust with respect to choice of distribution an experimenter assigns to these intervals.

It appears that whether selected intervals for rates of mutations are robust with respect to distributions assigned to them may also depend on the relative lengths of these intervals. For example, the ratio of length of the intervals for transitions in experiment VI to that used in experiments IV and V is

$$\frac{2 \times 10^{-7}}{10^{-6} - 10^{-7}} = 0.2222, \tag{8.6.2}$$

when truncated to four decimal places, and the same ratio for transversions is

$$\frac{2 \times 10^{-7.5}}{10^{-7} - 10^{-8}} = 0.7027, \tag{8.6.3}$$

when truncated to four decimal places. Thus, the lengths of the intervals of rates used in experiment VI were only fractions of those used in experiments

IV and V. It is also of interest to compare the lengths of the interval used for rates of transversions in experiments VI to that used for transitions in this experiment. This ratio is

$$\frac{2 \times 10^{-7.5}}{2 \times 10^{-7}} = 0.3162 \tag{8.6.4}$$

when truncated to four decimal places. Interestingly, the length of the interval for transversions is less than a third of that for transitions in experiment VI. These observations also suggest that the tightness of these intervals in experiment VI contributed the robustness with respect to choice of distribution of mutation rates for these intervals.

8.7 Potential Applications of Many Site Models to the Evolution of Protein Coding Genes

In section 8.5 attention was focused on the evolution of a three letter codon by the process of nucleotide substitution, but in reality, however, protein coding regions of DNA molecules usually consist of many three letter codons. In principle, therefore, applications of the model of nucleotide substitution at many sites under consideration is not limited to only problems that arose in connection with the Genographic Project but may also be applied to provide insights into the evolution of protein coding genes consisting of many three letter codons. Li (1997) in his chapter 7 has provided a valuable review of rates and patterns of nucleotide substitution for various mammalian and Drosophila protein coding genes.

According to Li, empirical rates of nucleotide substitution are estimated as follows. Let T denote the time, expressed in years, of divergence of two related species based on paleontological data. Suppose that two homologous sequences of DNA from the two species are aligned so that it is possible to count the number of sites where nucleotides differ for the two strands under consideration. Let K denote the number of sites where they differ. Then, rates of nucleotide substitution are estimated by the ratio

$$r = \frac{K}{2T}. \tag{8.7.1}$$

The rationale for the factor 2 in the denominator was not discussed by the author, but, evidently, this factor arises because two strands are being compared and investigators wish to express rates per strand per unit time. Interesting variations have been observed in rates of nucleotide substitutions as measured by the formula given above.

A nucleotide substitution in a three letter codon is said to be synonymous if a mutated codon codes for the same amino as the codon form which it mutated. By definition, if a nucleotide substitution in a codon that is not synonymous it is called nonsynonymous. Interestingly, observed rates of nonsynonymous mutations are lower than those for synonymous mutations for several mammalian genes, see table 7.1 of Li (1997). These differences in rates have been attributed to the action of natural selection, for, if a mutation is nonsynonymous, then the coding for protein, consisting of many amino acids, will be changed to another protein, which may be detrimental to the ability of an individual to survive and reproduce. A very useful extension of the software described in this chapter would be that of writing code to indicate whether a nucleotide substitution in a codon results in a synonymous or nonsynonymous mutation. Furthermore, for those proteins that have been sequenced, one could write code to record the changes in the amino acids making up the proteins attributable to the mutational process of nucleotide substitution.

Another direction in which it would be of interest to extend the stochastic model of nucleotide substitution at many sites under consideration would be that of accommodating natural selection in the formulation. A first step in this direction that would need to be taken into account would be of generation times of individuals in the species under consideration. These times consist of two components in populations consisting of two sexes, females and males. One component is the time following birth to the time that females and males are capable of producing offspring, and a second component would be the survival times following the birth of offspring. From an evolutionary perspective, this second component would be very significant, because the ability of parents to pass on their DNA to descendants would depend, in large part, on their ability to survive and raise their offspring to sexual maturity.

According to the model of nucleotide substitution at many sites of a DNA molecule under consideration, the waiting times among mutations are random variables expressed in years and based on paleontological data. To convert these times to generation times, let the random variable T denote the time among mutation in the multiple site model under consideration, and let T_R denote a random representing times, ages, females and males in a population reach sexual maturity. The time units of both these random variables could be years. Then, formally, $P\left[T \leq T_R\right]$ is the probability that an individual who has reached sexual maturity carries a mutation at some site of a DNA molecule in the set of sites under consideration. There is, of course, at least one caveat regarding this formulation; namely this probability may differ among females and males.

As a zygote develops from a one-celled entity to a mature adult, the most likely times for nucleotide substitutions to occur is in the process of cell division. Moreover, the greater the number of cell divisions in an individual, the greater is the likelihood that an individual is a carrier of a mutation in the form of a nucleotide substitution. Furthermore, if a mutation occurs during meiosis, the cell division leading to the production of gametes, then there is a positive probability that it will be transmitted to an offspring. Because males produce vast numbers of sperm but females produce relatively few eggs so that there are many more meiotic cell division in males than in females, it seems plausible that males will pass on a nucleotide substitution to their offspring with higher probability than that for females. Such chains of reasoning provide a rationale for the idea that the probability $P\left[T \leq T_R\right]$ is higher in males than in females. A problem that frequently arises is such discussion is that of converting paleontological time to cell division when considering the probability of observing a nucleotide substitution per meiotic division.

For further discussion of molecular evolution, it is recommended that the books, Li (1997), Nei and Kumar (2000) and Yang (2006) be consulted. The subject of nucleotide substitution is still an active field of investigation, and by entering the topic “nucleotide substitution” into a search engine for the world wide web one can find references to many current and recent research papers on various topics in the field.

8.8 Preliminary Notes on Stochastic Models of Indels and Other Mutations

The word indel is a portmanteau of two types of genetic mutations called insertions and deletions. Unlike nucleotide substitutions, which involve a change in only one nucleotide, indels involve changes in several nucleotides at a site or some set of sites in a molecule of DNA. Historically, the term seems to have arisen in studies of comparative genomics. Suppose in such a study, for example, that species A at some locus has 4 repeats of the nucleotide C and species B has 5 repeats of C at this locus. If the evidence suggests that no natural selection was involved in the evolution of this locus in both species, then it seems plausible that either species A evolved from species B by a deletion event of one nucleotide or species B evolved from species A by an insertion event of one nucleotide. When ambiguities of this type arise in the interpretation of such observations, the mutations are often referred to as indels.

If the word, indels, is entered into a search engine for the world wide web, then many web sites will appear on the computer monitor. Among sites that may be accessed from such searches are the effects of indels that are mutations referred to as "shifts the reading frame" in genes that code for proteins. Examples of such mutations occur when an indel creates a new stop codon so that the mutated genes codes for a different sequence of amino acids than that for the original gene. Other types of mutations involve translocations, in which DNA from one chromosome is inserted into another chromosome, and inversions, which are involved in the reversal of a set of nucleotides at some locus. The modelling of such mutations will not, however, be considered in detail in this section. In a recent paper, Levy *et al.* (2007), the diploid genome of an individual, J. Craig Venter, was presented. Briefly, in this diploid genome non-SNP variation accounted for 22% of all events identified and they were also involved in 74% of all variant bases. Such data point to a need for stochastic models that accommodate indels and other non-SNP types of mutation when considering problems that arise in molecular evolution.

Kimmel *et al.* (1996), and Kimmel *et al.* (1998) have used the following type of model in their studies of the dynamics of repeat loci in populations. For further information on these step-wise mutation models, the references in these papers may be consulted. In the models considered by Kimmel *et al.*, the number of repeats of a single nucleotide, pairs of nucleotides or sets of codons are considered. At some locus on a chromosome, let the random variable X denote the number of repeats of some structure in the DNA. To help fix ideas, suppose the nucleotide C occurs only once at some locus. Then, $X = 0$, but if the repeat CC is observed, then $X = 1$. Evidently, the range of the random variable X is some finite subset of the set of non-negative integers $(n \mid n = 0, 1, 2, \ldots)$. Alternatively, if one wished to consider the evolution of indels, then a random variable X could represent the number of nucleotides in some locus under consideration. In such cases, the range of the random variable X would be finite subset of the set of positive integers $(n \mid n = 1, 2, \ldots)$.

To set up a framework for a step-wise mutation model involving the evolution of indels, let U denote a random variable taking values in a finite subset of the set $(n \mid n = 0, \pm 1, \pm 2, \pm 3 \ldots)$ of all possible integers. Let X_0 denote the initial size of a locus under consideration expressed in terms of the number of nucleotides and consider a sequence of random variables $(X_k \mid k = 0, 1, 2, 3, \ldots)$, where X_k is the size of the locus after $k \geq 1$ generations. Next let $(U_k \mid k = 1, 2, 3, \ldots)$ be a sequence of independent random variables such that each random variable in the sequence has the same distribution as the random variable U. Then, in a step-wise mutation model, it is assumed that for every $k \geq 1$ the recursive relation

$$X_k = X_{k-1} + U_k \tag{8.8.1}$$

is satisfied. Observe that if $U_k = 0$, then no mutation occurs, but if $U_k \neq 0$, then a mutation occurs, In particular, if $U_k < 0$, then a deletion occurs involving of U_k nucleotides occurs, and if $U_k > 0$, then an insertion of U_k nucleotides occurs. Such insertions or deletions could be repeats of nucleotides or some other units of DNA such as codons.

In the process of constructing any step-wise mutation model in the class models under consideration, the choice of the distribution of the random variable U will be crucial. Among the many approaches to constructing this distribution, the following procedure is of interest, because of its simplicity. Suppose, for example, it is known that a mutation is a rare event and let $\eta \in (0, 1)$ denote the probability per generation that a mutation does not occur. Then,

$$P\left[U = 0\right] = \eta \tag{8.8.2}$$

and $1 - \eta$ is the probability of a mutation occurring per generation. Next consider a random variable V with a uniform distribution on the set of integers $(n \mid n = 1, 2, \ldots, m)$, where $m \geq 1$ is the maximum number of units of DNA that may be deleted or inserted. Let $V = v$ be a realization of the random variable V and let ξ denote a random variable such that $P\left[\xi = 1\right] = \zeta$ and $P\left[\xi = -1\right] = 1 - \zeta$, where $\zeta \in (0, 1)$. Then, given that $U \neq 0$ with probability η, a realization of the random variable $U = u$ would be computed using the equation $u = \xi v$, where $\xi = 1$ with probability ζ and $\xi = -1$ with probability $1 - \zeta$.

To obtain a better grasp of the properties of the three parameter model just described, it is appropriate to further discuss their interpretations. Because mutations are usually rare events, the parameter η will be chosen such that $1-\eta \in \left(10^{-7}, 10^{-4}\right)$ or some other interval that seems appropriate because of prior knowledge of probabilities of mutations per generation. Values of the parameter $\zeta \in (0, 1)$ govern the likelihood of insertions and deletions. If, for example, $\zeta = 0.5$, then insertions and deletions will be equally likely, if $0 < \zeta < 0.5$, then deletions are more likely than insertions and if $0.5 < \zeta < 1$, then insertions are more likely than deletions. If $\zeta = 1$, then only insertions will occur, and if $\zeta = 0$, then only deletions will occur.

To further fix ideas, suppose the sequence of random variables in (8.8.1) is viewed as a Markov chain with some finite state space

$$\mathfrak{S} = (x_1, x_2, \ldots), \tag{8.8.3}$$

where $0 \leq x_1 < x_2 < \ldots$ are integers. If, for example, $x_1 = 0$, then when the process enters this state, the locus under consideration is empty or disappears, but, in principle, in a subsequent step in the evolution of a DNA molecule, the locus may reappear when some nucleotides are inserted into this particular location. These observations suggest that the parameter m in generation $k \geq 2$ will depend on the state of the process in generation $k-1$. In particular, if $X_{k-1} = x \in \mathfrak{S}$, then suppose $m(x)$ is a number such that $1 \leq m(x) \leq x$. If, for example, $m(x) = x$, then a deletion of size x nucleotides may occur so that the locus becomes empty or the locus may be doubled in size if an insertion of size x occurs. For cases in which $m(x) \geq 1$ is much smaller than x, the evolution of changes in the size of a locus would proceed at a slower pace than if $m(x) = x$. Alternatively, the distribution of deletions and insertion may be asymmetric. In this situation, given that x is the size of the locus in generation $k-1$, let $m_D(x)$ denote the maximum number of deletions that may occur and let $m_I(x)$ denote the maximum number of insertions that may occur. Then, $1 \leq m_D(x) \leq x$. but $m_I(x)$ may be any number such that $m_I(x) \geq 1$.

If natural selection acts in such a way that it tends to conserve the size of a locus, then the process may evolve toward a Markov chain in which all states in $\mathfrak{S}$ communicate so that the process would converge to a stationary distribution. A population in a statistical equilibrium, governed by this stationary distribution, would consists of individuals with variable numbers of nucleotides with respect to the locus under consideration. An example of such a locus may be that for Huntington's disease (HD), which occurs in the short arm of chromosome 4 of the human genome. In this locus, the codon, CAG, which codes for the amino acid glutamine, is repeated a variable number of times among individuals. If an individual has some number $r \geq 1$ repeats of this codon, then chains of r glutamines, known as polyglutamines (*poly*(Q)), arise from this set of repeats. A *poly*(Q) of 36 glutamines produces a cytoplastic protein called huntingtin, symbolized by *Htt*, whereas one with 40 or more repeats produces mutant form of the protein *Htt* symbolized by *mHtt*. The expression of such mutant proteins varies among individuals as to ages of onset of physical and cognitive decline. In some individuals, the onset of HD occurs before the reproductive age so they can not reproduce, which are examples of cases where natural selection tends to eliminate HD genes with 40 or more repeats of the codon CAG from a population. Huntington's disease has been researched extensively so that a wealth of information on this disease may be obtained by entering this term into a search engine for the world wide web.

There are many non-SNP types of mutations that have not been mentioned in this section. Among these types are gene duplication, unequal cross-overs and nondisjunction of chromosomes during meiosis as well as insertions of DNA from another species into the genome of a species under consideration. For an extensive account of these and other mutations and their roles in evolution, the book Li (1997). may be consulted.

Another approach to designing models for indels is to let Y_1 and Y_2 denote two independent random variables taking values in the set of nonnegative integers $(y \mid y = 0, 1, 2, \ldots)$ with specified distributions. Then, the random variable $Y_1 - Y_2$, which takes values in the set of all integers $(y \mid y = 0, \pm 1, \pm 2, \ldots)$, could be used as a model for indels. In this formulation negative and positive values would represent deletions and insertions, respectively, and 0 would represent no mutations. For further discussions of this type of model the papers of Kimmel *et al.* cited above may be consulted.

Bibliography

[1] Behar, D.M. *et al.* (2007) The Genographic Project Public Participation Mitochondrial Data Base. PLoS Genetics **3**:1083–1095.

[2] Felsenstein, J. (2004) **Inferring Phylogenies.** Sinauer Associates Publishers, Sunderland, Massachusetts.

[3] Ingman, M. (2001) Mitochondrial DNA Clarifies Human Evolution. ActionBioscience.org, May 2001.

[4] Kimmel, M. and Chakraborty, R. (1996) Measures of Variation at DNA Repeat Loci Under a General Stepwise Mutation Model. Theoretical Population Biology **50**:345–367.

[5] Kimmel, M., Chakraborty, R., Stivers, D. and Deka, R. (1996) Dynamics of Repeat Polymorphisms Under Forward-Backward Mutation Model: Within and Between Population Variability at Microsatellite Loci. Genetics **143**:549–555.

[6] Kimmel, M., Chakraborty, R., King, J., Bamshad, M., Watkins, W. and Jorde, L.(1998). Signatures of Population Expansion in Microsatellite Repeat Data. Genetics **148**:1921-1930.

[7] Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halperin, A. L. *et al.* (2007) The Diploid Genome Sequence of an Individual Human. PLoS Biol. 5(10): e254 doi:10,1371/journal pbio 0050254.

[8] Li, W. H. (1997) **Molecular Evolution**. Sinauer Associates Inc. Sunderland, Mass 01375.

[9] Nei, M. and Kumar, S. (2000) **Molecular Evolution and Phylogenetics**. Oxford University Press.

[10] Thisted, R. A. (1988) **Elements of Statistical Computing-Numerical Computation**. Chapman Hall, New York and London.

[11] Wells, S. (2006) **Deep Ancestry - Inside the Genographic Project**. National Geographic Society, Washington, D. C., U.S.A.

[12] Yang, Z. (2006) **Computational Molecular Evolution**. Oxford University Press.

Chapter 9

Genealogies, Coalescence and Self-Regulating Branching Processes

9.1 Introduction

The subject of coalescence, which deals with properties of genealogies in humans and other species, is retrospective in the sense that, given a sample of individuals at some point in time, the objective is to draw inferences about its genealogical history. As an illustrative example, suppose we consider a random sample of size n from some human population at some point in time t. Each individual in this sample has two parents, his mother and father, who in turn had two parents. This process of enumeration could be continued back in time generation by generation and it can be seen that the number ancestors in each preceding generation increases as a power of 2. That is there are $2 = 2^1$ parents, $4 = 2^2$ grandparents, $8 = 2^3$ great grandparents and so on for every individual in the sample. If, therefore, one considers the number of ancestors that any individual in the sample has s generations back in his genealogy, it is easy to see that this number is 2^s. Even when s is not a large integer, 2^s can be a large number. For example, if $s = 50$, then, according to the symbolic calculation engine connected with this word processor,

$$2^{50} = 1,125,\ 899,\ 906,\ 842,\ 624.$$

It will be instructive to compare this number with the approximate size of the human population of the earth at the present time. For illustrative purposes, suppose it is six billion. According to American English, a billion is a thousand million. In symbols,

$$1000,000,000 = 1,000,000,000 = 10^9$$

So six billion is the number

$$6000,000,000 = 6,000,000,000 = 6 \times 10^9.$$

Now suppose we compute the fraction

$$\frac{6 \times 10^9}{2^{50}} = 5.329\,070\,518\,200\,75 \times 10^{-6}.$$

As this calculation illustrates, the total population of the earth at the present time is a small fraction of the number 2^{50}. According to some estimates, the carrying capacity of the earth is about 15×10^9 people. Even this larger number is a rather small fraction of the number 2^{50}.

$$\frac{15 \times 10^9}{2^{50}} = 1.332\,3 \times 10^{-5}$$

To get some idea of the number of years 50 generations may represent, suppose the time from birth to age of reproduction for the human female is in the range 15 to 30 years. Then, 50 generations of births would span the time interval of 750 to $1,500$ years. In terms of western civilization, 1500 years would take us back in time to the medieval period, and if, for example, a census of western Europe had been taken at that time, then the estimated population size would be much less than 2^{50}. A similar statement would be valid for any other geographical region of the earth. This simple illustration suggests that the current population of the earth consists of descendants of a much smaller population 1500 or more years ago. Furthermore, if one goes back $10,000$ years to the last ice age, the present population of the earth would be descendants of an even smaller population of ancestors. One reaches the conclusion, therefore, that the current population of the earth consists of a related set of individuals who have in common a rather small ancestral population. The DNA evidence supports this conclusion in that 98 to 99 percent of the base sequences of human beings tested match.

To return to the sample of size n, consider a thought experiment in which we trace the genealogy of each individual back in time. At some point in the preceding generations of the ancestral population of this sample, there would be a first generation such that at least two individuals would have at least one ancestor in common. When this event occurs, it is called the first coalescence of genealogies. By continuing in this way, there would be a second and third coalescence, and, in principle, if the records were available one could construct a complete set of genealogies and a coalescent for the $n > 1$ individuals in the sample. With the possible exception of some families or small nations, such records are not available for extended periods of time. Consequently, it becomes necessary to construct probabilistic models of coalescence in samples of either individuals, genes, cells or even biological entities such as viruses.

As is turns out, the formalization of the mathematics designed for human genealogies comprising two sexes leads to difficult conceptual and computational problems. Consequently, existing theories on coalescence are limited to haplotypes which may include genes. From the point of view of human populations, such theories apply to mitochondrial DNA inherited through maternal lines and some regions of the DNA on the Y chromosome inherited through paternal lines, where, during meiosis, there is no exchange of genetic material. Two very influential initial papers in the field of coalescence are those of Kingman (1982a) and (1982b). Briefly, Kingman worked within a paradigm of constant populations size from generation to generation and showed that as population size N becomes large, the coalescent process converges to a Markov chain in continuous time whose state space is a set of equivalence classes and the time unit is expressed in terms of units of length N. There is an extensive literature on generalizations of Kingman's ideas and a review of the literature may be found in the paper of Nordborg (2001). An interested reader may also find accounts of more recent ideas about coalescence by typing, coalescence genetics into a search engine for the world wide web. Originally, the idea of coalescence was conceived of in terms of genes so that waiting times among coalescent events would measure in the millions of years. But, recently, it appears some authors are thinking about shorter evolutionary time periods when applying the theory. More recent contributions to this field include those of Jagers and Sagitov (2004) and Pollak (2007). The paper of Jagers and Sagitov was devoted to a case in which population size may vary among generations but in large populations there is still convergence to the coalescent. In Pollak (2007), coalescence for a completely random mating population is considered. An overview of the subject of coalescence may be found in chapter 7 of the book Haccou, Jagers and Vatutin (2007). Extensive applications of the ideas embodied in coalescence theory may be found in the recent book Durrett (2008) in which attention is confined to Wright-Fisher processes and other processes characterized by the condition that population size is constant from generation to generation. A distinguishing feature of this book is that many results are stated as theorems followed by proof, which will be welcomed by many mathematically inclined readers.

In this chapter, however, the subject of coalescence will be approached within a theory of self-regulating branching process by developing structures within which it is actually possible to simulate genealogies stemming from one or more initial individuals. By using this computer intensive approach, it will be possible, in principle, to accommodate not only population sizes that vary among generations but also various types of mutation and selection. But, because the computer implementation of these ideas will give rise to a need for an extensive development of software, only some theory underlying computing algorithms will be outlined in this chapter along with a few illustrative computer experiments. Unlike Kingman type theories of coalescence in which one has a retrospective perspective, in the approach to be illustrated in this chapter, one simulates a genealogy by going forward in time and then looks back in time by sampling randomly chosen pairs of individuals in some generation and finding the generation back in time where they had a common ancestor. When designing these models and simulation experiments, rather short periods of evolution of about 100,000 to 200,000 years were under consideration.

9.2 Simulation of Stochastic Genealogies – One Type Case

Most groups of people, who have lived in some geographical area for a long period of time, have stories regarding their origins. These stories quite often refer to a population of founders from which all individuals in the population descend. Usually the founder population consists of one individual or a set of individuals. For the sake of simplicity, attention will initially be focused on the case in which all individuals in a population are descended from one initial individual denoted by (0). The first step in formalizing the idea of stochastic genealogies is that of developing a notation that accounts for every individual in the set of descendants of the initial individual (0). As the discussion progresses, attention will also be given to the problem of creating data structures in a computer containing all descendants of (0) up to some generation $n \geq 1$ for a sample of $N \geq 1$ realizations of a stochastic genealogy process. Attention will also be given to arranging arrays in a computer such that it will be possible to follow a line of descent from any individual in generation n back in time to the initial individual (0) or forward in time to some individual in generation $n \geq 1$. Let the symbol $\Omega_n = (\omega)$ denote the set of all realizations of the process up to $n \geq 1$ generations, where every $\omega \in \Omega_n$ has the form $\omega = (\omega_0, \omega_1, \omega_2, \ldots, \omega_n)$. The $\omega's$ are abstract symbols formalizing the notion of variability among realizations of the process. A realization of a stochastic genealogy for n generations will be sets of individuals in generation $1 \leq \nu < n-1$ denoted by $\mathcal{I}_\nu(\omega_\nu)$, for $\nu = 0, 1, 2, \ldots, n$, such that individuals in the set $\mathcal{I}_\nu\left(\omega_\nu\right)$ of individuals in generation ν are the offspring of the individuals in the set $\mathcal{I}_{\nu-1}\left(\omega_{\nu-1}\right)$ in generation $\nu-1$ and $\mathcal{I}_0\left(\omega_0\right) = (< 0 >)$. As the points $\omega = (\omega_\nu \mid \nu = 0, 1, 2, \ldots, n)$ vary over the set Ω_n so does the set individuals in a genealogy

$$\mathcal{I}\left(\omega\right) = \bigcup_{\nu=0}^{n} \mathcal{I}_\nu\left(\omega_\nu\right). \tag{9.2.1}$$

As a first step in defining a stochastic mechanism characterizing the variation among samples $\omega \in \Omega_n$ of genealogies, it will be supposed that the number of offspring contributed by the initial individual to generation 1 is a realization of a random variable X taking values in the set $(x \mid x = 0, 1, 2, \ldots)$ of nonnegative integers. Let x_0 be a realization of the random variable X and let the symbols $(0, \nu_1)$, for $\nu_1 = 1, 2, \ldots, x_0$, denote the individuals of the generation 1. If $x_0 = 0$, then the process stops with the initial individual (0). But, for $x_0 \geq 1$, the set of individuals of generation 1 will be denoted by array

$$\mathcal{I}_1\left(\omega_1\right) = \left((0, \nu_1) \mid \nu_1 = 1, 2, \ldots, x_0\right), \tag{9.2.2}$$

which are the x_0 offspring of (0). In computer representation of the set $\mathcal{I}_1\left(\omega_1\right)$, individuals would be ordered in the natural order $\nu_1 = 1, 2, \ldots, x_0$ and $\mathcal{I}_1\left(\omega_1\right)$ could be shaped into a $x_0 \times 2$ array.

To continue the construction of a stochastic genealogy, it will be assumed that, given $X = x_0 \geq 1$, all individuals of generation 1 produce offspring independently according to the random variables $X_{\nu_1}, \nu_1 = 1, 2, \ldots, x_0$, which are independently and identically distributed copies of the random variable X. Let $X_{\nu_1} = x_{\nu_1}$, for $\nu_1 = 1, 2, \ldots, x_0$, be realizations of these random variables and let the symbol $(0, \nu_1, \nu_2)$ denote the ν_2-*th* offspring in generation 2 of the individual $(0, \nu_1)$ in generation 1. Symbolically, the set $\mathcal{I}_2\left(\omega_2\right)$ of realized individuals in generation 2 may be represented by the disjoint union singletons

$$\mathcal{I}_2\left(\omega_2\right) = \bigcup_{\nu_1, \nu_2} ((0, \nu_1, \nu_2)), \tag{9.2.3}$$

where $\nu_1 = 1, 2, \ldots, x_0$ and, for each ν_1, $\nu_2 = 1, 2, \ldots, x_{\nu_1}$. If $x_{\nu_1} = 0$ for some ν_1, then individual $< 0, \nu_1 >$ of generation 1 does not contribute any offspring to generation 2 and, therefore, has no descendants in a genealogy. The set $\mathcal{I}_2\left(\omega_2\right)$ of individuals in generation 2 may be ordered in many ways, depending on the way a set of simulated genealogies will be used. In what follows, emphasis will be placed on developing an ordering procedure such that the most recent common ancestor of any two individuals in some generation $n \geq 2$ in a simulated genealogy may be unambiguously identified in some generation $n_0 < n$. As a first step in developing this system of ordering, it will be helpful to consider some specific examples to illustrate how each individual in a genealogy may be labeled.

As stated above, suppose a population evolves from one individual denoted by the 1×1 array (0). Then suppose that the initial individual has two offspring, and let

$$\mathcal{I}_1(\omega_1) = \begin{pmatrix} 0 & 1 \\ 0 & 2 \end{pmatrix} \tag{9.2.4}$$

denote a 2×2 array representing the of two individuals of generation 1. Each row of this array describes a line of decent from the initial individual and is a distinct label for each individual. When constructing such an array in a computer, one way of accomplishing the construction in (9.2.4) is to create the string $(0, 1, 0, 2)$ and then shape into a 2×2 array.

Next suppose each individual in this array has two offspring. Then the set of lines of descent from the initial individual (0) may be represented in the 4×3 ordered array

$$\mathcal{I}_2(\omega_2) = \begin{pmatrix} 0 & 1 & 3 \\ 0 & 1 & 4 \\ 0 & 2 & 5 \\ 0 & 2 & 6 \end{pmatrix}. \tag{9.2.5}$$

A useful way of constructing this array in a computer it to construct a string of length 12 and then shape into a 4×3 array. Each individual in this array has a distinct label that distinguishes this individual from any other individual in the genealogy. From this array, it is easy to find the most recent ancestor of any two selected individuals. For example, if the individuals $(0, 1, 3)$ and $(0, 1, 4)$ are selected, then the individual $(0, 1)$ in generation 1 is the most recent ancestor of these two individuals. But, if the individuals $(0, 1, 3)$ and $(0, 2, 5)$ are selected, then the initial individual (0) is the most recent ancestor of the two individuals.

Next suppose each of the four individuals in this array has two offspring. Then the eight individuals in generation 3 may be represented by the 8×4 ordered array

$$\mathcal{I}_3\left(\omega_3\right)=\begin{pmatrix}0 & 1 & 1 & 7\\ 0 & 1 & 1 & 8\\ 0 & 1 & 2 & 9\\ 0 & 1 & 2 & 10\\ 0 & 2 & 3 & 11\\ 0 & 2 & 3 & 12\\ 0 & 2 & 4 & 13\\ 0 & 2 & 4 & 14\end{pmatrix}. \tag{9.2.6}$$

Again, each individual in this array has the label that distinguishes it from any other individual in a genealogy. If the two individuals $(0,1,1,7)$ and $(0,1,1,8)$ in generation 3 are chosen, then it can be seen that the individual $(0,1,1)$ in generation 2 is the most recent common ancestor, and if the two individuals $(0,1,1,7)$ and $(0,1,2,9)$ of generation 3 are chosen, then it can be seen that the individual $(0,1)$ in generation 1 is the most recent ancestor of both these individuals. Similarly, if the individuals $(0,1,1,7)$ and $(0,2,3,11)$ of generation 3 are chosen, then it can be seen that the initial individual (0) is the most recent ancestor of these individuals.

In general, let $\iota=(\nu_0,\nu_1,\nu_2,\nu_3)$ and $\iota'=(\nu_0',\nu_1',\nu_2',\nu_3')$ denote two arbitrary distinct individuals in generation 3. Then, by using the observations outlined above, it can be seen that if $\nu_0=\nu_0'$ but $\nu_k\neq\nu_k'$ for $k=1,2,3$, then $(\nu_0=0)$ is the most recent common ancestor of these two individuals. And if $\nu_k=\nu_k'$ for $k=0,1$ but $\nu_k\neq\nu_k'$ for $k=2,3$, then the individual (ν_0,ν_1) in generation 1 is the most recent ancestor of these two individuals. By continuing in this way, it can be seen that for any two distinct individuals in generation 3, their most recent common ancestor could be determined. The genealogies of two individuals are said to coalesce at their most recent common ancestor. Our next task is to generalize these observations in such a way that for any two distinct individuals ι and ι' in some generation $n\geq 1$, it will be possible to write computer code to identify the most recent common ancestor of these individuals. The next step in developing a system of ordering individuals and labeling individuals in successive generations is to generalize these observations to include the cases in which each individual in some generation contributes a random number of offspring to the next generation.

To see how to order the individual of generation 2 in the general case, let $x_0\geq 1$ and suppose as above that individual $(0,\nu_1)$ has x_{ν_1} offspring for $\nu_1=1,2,3,\ldots,x_0$. For $\nu=0,1,2$, let y_ν denote the number of individuals in generation ν. Then, $y_0 = 1, y_1 = x_0$ and y_2 is given by the sum

$$y_2 = \sum_{\nu_1=1}^{x_0} x_{\nu_1}. \tag{9.2.7}$$

In a computer, the set $\mathcal{I}_2(\omega_2)$ of individuals for generation 2 could be shaped into a $y_2 \times 3$ array. To construct this array in a computer imagine a string of length $y_2 \times 3$ partitioned into sub-strings of length 3. The first sub-string would consist of the labels $((0,1,x_0+1),(0,1,x_0+2),(0,1,x_0+3))$, the next sub-string would consist of the labels $((0,1,x_0+4),(0,1,x_0+5),(0,1,x_0+6))$ and so on down to the set of labels

$$((0,1,x_0-(x_1-2)),(0,1,x_0+(x_1-1)),(0,1,x_0+x_1)), \tag{9.2.8}$$

which would complete the listing of offspring for the first individual $(0,1)$ in the set of individuals $\mathcal{I}_1(\omega_1) = ((0,\nu_1) \mid \nu_1 = 1,2,\ldots,x_0)$ for generation 1. The next step in creating a string of length $y_1 \times 3$ would be that of labeling all offspring of the second individual in row 2 of the set $\mathcal{I}_1(\omega_1)$ in sub-strings of three. It is clear that this process could be continued until all offspring of the individuals were labeled in a string of $y_1 \times 3$ elements. Then to construct the set $\mathcal{I}_2(\omega_2)$ this string could be shaped into a $y_2 \times 3$ array. Of course, if $x_{\nu_1} = 0$ for some ν_1, then this individual in generation 1 would contribute no offspring to generation 2. It is also clear that this process could be continued recursively to construct a realization of a genealogy up to $n \geq 2$ generations such that each individual has a unique label.

Now suppose this process of constructing a genealogy is continued up to generation $n-1$. Then, in this generation, the set $\mathcal{I}_{n-1}(\omega_{n-1})$ would consist of individuals of the form $\iota = (\nu_0, \nu_1, \nu_2, \ldots, \nu_{n-1})$, where $\nu_0 = 0$, and the elements of ι are ordered and labeled into a $y_{n-1} \times n$ array, where ν_{n-1} is a distinct integer which identifies this individual uniquely in generation $n-1$. For every $\iota \in \mathcal{I}_{n-1}(\omega_{n-1})$, let x_ι be a realization of a random variable X_ι denoting the number of offspring individual ι contributes to generation n. Given $y_{n-1} \geq 1$, it will be assumed that the random variables in the collection

$$(X_\iota \mid \iota \in \mathcal{I}_{n-1}(\omega_{n-1})) \tag{9.2.9}$$

are independent copies of the random variable X. Then, the total number of individuals in generation n is given by the sum

$$y_n = \sum_\iota x_\iota, \tag{9.2.10}$$

where the index ι runs over the set $\mathcal{I}_{n-1}(\omega_{n-1})$ and, altogether, there are y_{n-1} terms in this sum.

The y_n individuals in the set of individuals $\mathcal{I}_n(\omega_n)$ of generation n will be denoted by the vectors $\iota = (\nu_0, \nu_1, \nu_2, \ldots, \nu_{n-1,}\nu_n)$ and arranged in a $y_n \times (n+1)$ array. Let ι_ν, for $\nu = 1, 2, \ldots, y_n$, denote the ordering of this array as described above. Then, the first x_1 rows in this array represent the offspring of individual 1 in generation $n-1$, the second set of rows represent the x_2 offspring of individual 2 in generation $n-1$ and so the ordering process continues until all the y_n individuals of generation n are listed in the set $\mathcal{I}_n(\omega_n)$. If $x_\nu = 0$ for some $\nu = 1, 2, \ldots, y_n$, then the individual $\iota_\nu = (\nu_0, \nu_1, \ldots, \nu_{n-1})$ in generation $n-1$ contributes no descendants to generation n. Suppose the software being used to develop the ideas just sketched has the property that if the symbols $\iota = (\nu_0, \nu_1, \ldots, \nu_n)$ and $\iota' = (\nu_0', \nu_1', \ldots, \nu_n')$ in the set $\mathcal{I}_n(\omega_n)$ of individuals in generation n are selected at random and equated $\iota = \iota'$, then the result is random Boolean vector $\boldsymbol{\xi} = (\xi_0, \xi_1, \ldots, \xi_n)$ such that $\xi_k = 1$ if $\nu_k = \nu_k'$ and $\xi_k = 0$ if $\nu_k \neq \nu_k'$ for $k = 0, 1, 2, \ldots, n$. Then, the range of the random sum

$$N_0 = \sum_{k=0}^{n} \xi_k \tag{9.2.11}$$

is the set of integers $(n_0 \mid n_0 = 1, 2, 3 \ldots, n-1)$, because for every pair (ν_n, ν_n') has the property that $\nu_n \neq \nu_n'$ so that $\xi_n = 0$ for any two individuals in generation n. In particular, if $N_0 = n_0 \geq 0$ is a realized value of the random variable N_0, then it follows from the ordering and the labeling of the individuals in generation n, that $\nu_k = \nu_k'$ for $k = 0, 1, 2, \ldots, n_0$ and $\nu_k \neq \nu_k'$ for $k = n_0 + 1, \ldots, n$ so that the individual $(\nu_0, \nu_1, \ldots, \nu_{n_0})$ is the most recent ancestor the individuals ι and ι' have in common. Let $B = n - N_0$ denote a random variable giving the number of generations back in time for one to find the most recent ancestor of two randomly chosen individuals ι and ι' in generation n. Upon a little reflection, it can be seen that the range of the random variable B is the set of integers $(b \mid b = 1, 2, \ldots, n)$. Let $P[B = b] = h(b)$, for $b = 1, 2, \ldots, n$, denote the density function of the random variable B. By resampling a set of genealogies in generation n, one could get an estimate of this density. For a fixed sample $\mathcal{I}_n(\omega_{in})$ among the $i = 1, 2, \ldots, N$ samples, one could continue to sample two individuals taken two at time at random without replacement until a sample $(b_1, b_2, \ldots, b_m)$, $m \geq 1$, of values back in time to the most recent ancestor was obtained. Given this sample, one could collate the data to produce a histogram, representing the number m_k of times an integer $k = 1, 2, \ldots, n$ occurred in the sample. Then, by computing the ratio m_k/m, for $k = 1, 2, \ldots .m$, one could obtain an empirical estimate of the frequency distribution of a random variable B, representing the number of generations back in time that one would have to go to find the most recent ancestor of two randomly chosen individuals in the set $\mathcal{I}_n\left(\omega_i\right)$. As the estimated distribution of the random variable B may depend on y_i, the number of individuals in the set $\mathcal{I}_n\left(\omega_{in}\right)$, which will vary among realizations of the process, an investigator may wish to continue the sampling process until a sample of size m is obtained from each of the N samples mentioned above. This set of samples could then be pooled to obtain a combined estimate of the frequency distribution of the random variable B. The sequence of random variables $(Y_0, Y_1, \ldots)$ represents the random numbers of individuals in each generation that would be computed in simulating any particular array $\mathcal{I}_n\left(\omega_{in}\right)$ of a sample of genealogies $\left(\mathcal{I}_n\left(\omega_{in}\right) \mid i = 1, 2, \ldots, N\right)$. In the next section, an overview of the branching process that will be used in computing realizations of this sequence of random variables will be given. If an array manipulating language such as *APL* were used to implement the ideas outlined in this section, then the task of writing the computer code could be accomplished with relative ease.

9.3 Overview of the Galton-Watson Branching Process

Let $\left(\mathcal{I}\left(\omega_i\right) \mid i = 1, 2, \ldots, N\right)$ be a sample of size $N \geq 1$ of genealogies defined by the union in (9.2.1). Then, for each sample, the sizes of the generations for $n \geq 1, Y_{i\nu}$ for $\nu = 0, 1, 2, \ldots, n$, are realizations of a sequence of random variables $(Y_0, Y_1, Y_2, \ldots)$ taking values in the set of nonnegative integers and representing the size Y_n of generation $n = 0, 1, 2, \ldots$. In the construction under consideration, $Y_0 = 1$ with probability one and the other random variables in the sequence may be computed recursively as random sums of random variables. In general, given that $Y_{n-1} = y_{n-1}$, let X_ν for $\nu - 1, 2, \ldots, y_{n-1}$ denote independent copies of the random variable X. Then for $n \geq 1$, the random variable Y_n is the random sum

$$Y_n = \sum_{\nu=1}^{Y_{n-1}} X_\nu. \tag{9.3.1}$$

As a first step in developing explicit formulas, let $f(x)$ denote the probability density function of the random variable X defined for $x = 0, 1, 2, \ldots$. A useful choice for this density would be that for the Poisson distribution with the formula

$$f(x) = e^{-\lambda} \frac{\lambda^x}{x!}, \tag{9.3.2}$$

where $\lambda > 0$ is a parameter. As is well known, for this case the expectation of the random variable X is $E[X] = \lambda$. To deduce a formula for the expectation $E[Y_n]$, note that $E[Y_1] = E[X_1] = \lambda$, and

$$Y_2 = \sum_{\nu=1}^{Y_1} X_\nu. \tag{9.3.3}$$

Therefore,

$$E[Y_2 \mid Y_1 = X_1] = X_1 \lambda \tag{9.3.4}$$

so that

$$E[Y_2] = E[E[Y_2 \mid Y_1 = X_1]] = \lambda^2. \tag{9.3.5}$$

By an induction argument, it can be shown that for every $n \geq 1$,

$$E[Y_n] = \lambda^n. \tag{9.3.6}$$

Actually, this formula holds for any distribution of the random variable X such that $E[X] = \lambda < \infty$. From this formula, it can be seen that if $0 < \lambda < 1$, then $E[Y_n] \downarrow 0$ as $n \uparrow \infty$, but, if $\lambda = 1$, then $E[Y_n] = 1$ for all $n \geq 0$. But, for $\lambda > 1$, it follows that, $E[Y_n], n \geq 0$, is an increasing sequence. Thus, when one wishes to simulate a sample of realizations of genealogies, choosing an offspring distribution such that $\lambda > 1$ would produce an interesting set of genealogies.

If X has a distribution with expectation λ and variance σ^2, it can be shown that if $\lambda \neq 1$, then

$$var[Y_n] = \frac{\sigma^2 \lambda^n (\lambda^n - 1)}{\lambda(\lambda - 1)}, \tag{9.3.7}$$

but if $\lambda = 1$, then

$$var[Y_n] = n\sigma^2. \tag{9.3.8}$$

Interestingly, if $0 < \lambda < 1$, then $var[Y_n] \downarrow 0$ as $n \uparrow \infty$, but if $\lambda \geq 1$, then $var[Y_n] \uparrow \infty$ as $n \uparrow \infty$. For the case X has a Poisson distribution with parameter λ, $\sigma^2 = var[X] = \lambda$. For a deeper description of the process, which will be useful in designing simulation experiments, it will be helpful to derive a formula for the distribution of the random variable Y_n for every

$n \geq 1$. Perhaps the most succinct way of describing these distributions is to use the generating function of the random variable X, which by definition is

$$g(s) = E\left[s^X\right] = \sum_{x=0}^{\infty} f(x)\, s^x \tag{9.3.9}$$

for $0 \leq s \leq 1$. In particular, if X has a Poisson distribution with parameter $\lambda > 0$, then, as was shown in a previous chapter,

$$g(s) = e^{\lambda(s-1)}. \tag{9.3.10}$$

For $n \geq 1$, let $g_n(s)$ denote the generating function of the random variable Y_n. Because $Y_0 = 1$, it follows that $Y_1 = X$ so that $g_1(s) = g(s)$. Given that $Y_1 = y_1$, the random variable Y_2 is given by

$$Y_2 = \sum_{\nu=1}^{y_1} X_\nu, \tag{9.3.11}$$

where X_ν for $\nu = 1, 2, \ldots, y_1$ are independent and identically distributed copies of the random variable X. Therefore, the conditional generating function of the random variable Y_2, given that $Y_1 = y_1$, is

$$E\left[s^{Y_2} \mid Y_1 = y_1\right] = (g(s))^{y_1}. \tag{9.3.12}$$

Hence, the unconditional generating Y_2 is

$$g_2(s) = \sum_{y_1=0}^{\infty} f(y_1)\,(g(s))^{y_1} = g(g(s)). \tag{9.3.13}$$

By an induction argument, it can be shown that for every $n \geq 1$

$$g_{n+1}(s) = g_n\left(g(s) = g(g_n(s))\right. . \tag{9.3.14}$$

As was mentioned in the foregoing discussion, all lines descending from the initial individual (0) may become extinct. That is, $Y_n = 0$ for some $n \geq 1$. Let $(\Omega, \mathfrak{A}, P)$ denote the probability space underlying the process and let $[Y_n = 0]$ denote the set of $\omega \in \Omega$ such that $Y_n(\omega) = 0$. If $Y_n = 0$, then $Y_{n+1} = 0$ so that $[Y_n = 0] \subset [Y_{n+1} = 0]$ for all $n \geq 1$. Thus, the event that the process is extinct by generation $n \geq 1$ is given by the union

$$E_n = \bigcup_{\nu=1}^{n} [Y_\nu = 0] = [Y_n = 0], \tag{9.3.15}$$

and the probability of eventual extinction is

$$q = \lim_{n\to\infty} P[E_n] = \lim_{n\to\infty} P[Y_n = 0] = \lim_{n\to\infty} g_n(0). \tag{9.3.16}$$

To find a method for calculating q, let $q_n = P\left[Y_n = 0\right] = g_n\left(0\right)$. Then it follows that $\left(q_n\right)$ is an increasing sequence with q as the limit. From equation (9.2.12), it also follows that $q_{n+1} = g\left(q_n\right)$, and by letting $n \to \infty$, it can be seen that q satisfies the equation

$$q = g\left(q\right). \tag{9.3.17}$$

The branching process under consideration is known as the Galton-Watson process, and, a historical account of this process may be found in the book of Harris (1963) as well as many results concerning this process that will not be discussed here.

Also contained in this book is a theorem stating that if $E\left[X\right] = \lambda \leq 1$, then $q = 1$, but, if $\lambda > 1$, then q is the unique non-negative root in the interval $\left(0,1\right)$ of the equation $s = g\left(s\right)$. For the case that X follows a Poisson distribution, this equation takes the explicit form

$$s = e^{\lambda(s-1)}. \tag{9.3.18}$$

Thus, for this case, to calculate q, let $q_1 = \exp\left(-\lambda\right)$ for a chosen value of λ, then calculate the sequence $\left(q_n\right)$ recursively from the equation

$$q_{n+1} = \exp\left(\lambda\left(q_n - 1\right)\right) \tag{9.3.19}$$

for $n \geq 1$ Usually, the sequence $\left(q_n\right)$ will converge to q after a few iterations.

When designing an experiment to simulate a sample of genealogies up to n generations, the ideas that have just been discussed have useful applications. For example, if one supposes that X has a Poisson distribution with parameter $\lambda > 0$ and $\lambda = 2$, a call to the calculation engine yields the result

$$q = 0.203187869\,979980 \tag{9.3.20}$$

as the solution to the equation

$$q = \exp\left(2 \times \left(q - 1\right)\right) \tag{9.3.21}$$

such that $q < 1$. Therefore, a value of $\lambda = 2$ may be an interesting choice for λ, but for this choice of λ, the expected population size in generation n is

$$E\left[Y_n\right] = 2^n, \tag{9.3.22}$$

which becomes very large even for moderate values of n. It is also interesting to note that the formula in (9.3.7) for the variance of Y_n becomes large at a faster rate than does that for $E\left[Y_n\right]$. These observations suggest that to avoid simulating genealogy whose size would exceed the memory available in a computer, the number of generations n considered in an experiment would need to be restricted. For example, if $n = 20$, then

$$E\left[Y_{20}\right] = 2^{20} = 1,048,\ 576. \tag{9.3.23}$$

As of 2008, there are desktop computers on the market with memories of three gigabytes. In such a computer, it would be feasible to store an array of size $1,048,\ 576 \times 21$, but when one inspects the formula for the $var\left[Y_2\right]$, it can be seen that this number will be very large so that even for $n = 20$, one needs to proceed with caution when designing a simulation experiments.

Of course, in preliminary illustrative experiments, an investigator is free to choose the distribution of the random variable X. One such simple distribution is that determined by the generating function

$$g\left(s\right) = 0.05 + 0.75s + 0.20s^2. \tag{9.3.24}$$

For this distribution, the expected value of X is $E\left[X\right] = \lambda = 1.15$, its variance is $var\left[X\right] = 0.227\,5$ and the probability of extinction is $q = 0.25$. Because, the probability that $X = 1$ is 0.75 and $\lambda = 1.15$, one would expect that the population size would grow at a slower rate. In this case, the expected size of the population in generation $n = 100$ would be

$$E\left[Y_{100}\right] = \left(1.15\right)^{100} = 1,174,313.\,450\,700\,29. \tag{9.3.25}$$

It will be left as an exercise for the reader to calculate the value of $var\left[Y_{100}\right]$ and make assessments as to the feasibility of simulating genealogies for $n = 100$ generations on some available computer platform. The two examples discussed in this section are just two cases among countless numbers of cases that could be considered, but the discussion of other models will be deferred to other sections in this and later chapters.

The exercise of computing the sample of genealogies could also be expedited if it could be arranged that none in the sets in the sample were empty, indicating that, in some realizations of the process, the population had become extinct after $n \geq 1$ generations of evolution. There is a simple way of avoiding the possibility that in some realizations of the process the population becomes extinct. Suppose, for example, that the random variable X, which denotes the random number of offspring produced by each individual, has a Poisson distribution with a parameter $\lambda > 0$. Then, when the range of the random variable X is the set of nonnegative integers $(x \mid x = 0, 1, 2, \ldots)$, the range of the random variable $W = 1 + X$ is the set of positive integers $(w \mid w = 1, 2, \ldots)$, and the density function of the random variable is

$$g_W\left(w\right) = \exp\left(-\lambda\right)\frac{\lambda^{w-1}}{(w-1)!} \tag{9.3.26}$$

for $w = 1, 2, \ldots$. It can be shown that $E[W] = 1 + \lambda = \gamma$, $var[W] = var[X] = \lambda$ and the probability of extinction is $q = 0$. Let the sequence $(1, Y_1, Y_2, \ldots)$ denote the sizes of the generation of a Galton-Watson branching process generated by the independent copies of the random variable W. Then, because $\gamma = 1 + \lambda > 1$, in generation $n \geq 1$ the variance of the random variable Y_n is

$$var[Y_n] = \frac{\lambda\gamma^n(\gamma^n - 1)}{\gamma(\gamma - 1)}, \tag{9.3.27}$$

see formula (9.3.7). Therefore, if λ is chosen such that $0 < \lambda < 1$, then the $var[Y_n]$ could be made sufficiently small so that all the sets in the sample in a sample of genealogies could be stored in a desk top computer. The artificial device just described may be useful in conducting preliminary computer experiments, but it would be of more interest, form a biological point of view, if a self-regulating branching process could be defined such that the growth of a population starting from an initial individual would not grow without bound. In the next section, such a process will be described and analyzed.

Before preceding to the next section, however, it will be helpful to identify the class of stochastic processes to which the Galton-Watson process $(Y_0, Y_1, Y_2, \ldots)$ belongs. First of all note that the state space of the process is the set $\mathfrak{S} = (i \mid i = 0, 1, 2, \ldots)$ of nonnegative integers and the state 0, indicating that the population is extinct, is an absorbing state. Given that $Y_{n-1} = i \geq 1$, the conditional generating function of the random variable Y_n may be expressed in the form

$$g_n(s \mid i) = g^i(s) = \sum_{j=0}^{\infty} p_{ij} s^j. \tag{9.3.28}$$

Because the expression on the right depends only on $i \geq 1$, the Galton-Watson process may be formulated as a Markov chain with the stationary transition probabilities given by

$$P[Y_n = j \mid Y_{n-1} = i] = p_{ij} \tag{9.3.29}$$

for $n \geq 1$, where p_{ij} is the coefficient of s^i in the series expansion of $g^i(s)$. If $i = 0$, then $p_{ii} = 1$ and $p_{ij} = 0$ if $i \neq j$. Unlike the Markov chains studied heretofore, the state space $\mathfrak{S}$ for this class of processes is infinite.

Readers who are interested in pursuing the subject of branching processes in greater detail may consult the books Asmussen and Hering (1983), Athreya and Ney (1972) and Jagers (1975).

9.4 A Self-Regulating Galton-Watson Branching Process

Consider a population of $y \geq 1$ individuals in some generation. A basic component of a self-regulating branching process is the distribution of the number of offspring produced by each of the individuals who survive to reproduce the next generation. For a population of size y, let $p(y)$ be the probability an individual survives to reproduce in the next generation, let N be a random variable taking values in the set of non-negative integers $(y \mid y = 0, 1, 2, \ldots)$, representing the number of offspring produced by any individual per generation, and let $(\xi_k(y)), k = 1, 2, \ldots$, denote a sequence of independent Bernoulli random variables with common expectation $p(y)$ that are conditionally independent of the random variable N. Then, the number of offspring of any individual who survives to reproduce in the next generation is given by the random sum

$$\zeta(y) = \sum_{k=1}^{N} \xi_k(y), \tag{9.4.1}$$

where $\zeta(y) = 0$ if $N = 0$. Let the random variable $Y_n, n = 0, 1, 2, \ldots$ represent total population size in generation n and, given $Y_n = y \geq 1$, let $(\zeta_{nk}(y))$ be a sequence of conditionally independent and identically distributed random variables whose common distribution is that of ζ. Then, a density dependent Galton-Watson process is a Markov chain $(Y_n), n = 1, 2, \ldots$ whose state space is the nonnegative integers and with stationary transition probabilities determined by the recursive random sums

$$Y_n = \sum_{k=1}^{Y_{n-1}} \zeta_{nk}(Y_{n-1}), \tag{9.4.2}$$

for $n =, 1, 2, \ldots$. Again if $Y_n = 0$, then $Y_{n+1} = 0$ so that 0 is an absorbing state, indicating the population is extinct

Whenever an investigator utilizes computer intensive methods, it often becomes necessary to assume some explicit form of the distribution of the random variable N. Accordingly, it will be assumed that N has a Poisson distribution with positive parameter λ. Under this assumption, it follows that for $y_1 \geq 1$ the transition matrix of the Markov chain takes the form

$$P\left[Y_{n+1} = y \mid Y_n = y_1\right] = p_{y_1 y} = \exp\left(-y_1 \lambda p\left(y_1\right)\right) \frac{(y_1 \lambda p(y_1))^y}{y!}, \tag{9.4.3}$$

for $y = 0, 1, 2, \ldots$. If $y_1 = 0$, then $p_{00} = 1$ and $p_{0y} = 0$ for $y \geq 1$. Therefore, it follows that for $y_1 \geq 1$

$$E\left[Y_{n+1} \mid Y_n = y_1\right] = y_1 \lambda p(y_1) \quad forx \geq 1. \tag{9.4.4}$$

It is of interesting to observe in passing that the expression of the right in (9.4.4) does not depend on the Poisson distribution. For, if the distribution of N has the finite expectation $\lambda = E\left[N\right]$, then the equation remains valid. Consequently, many of the mathematical results that follow do not depend on the assumption of a Poisson distribution.

As, in general, $p(y)$ will be a nonlinear function of y, the unconditional expectation function, $E\left[Y_{n+1}\right]$, will not satisfy a linear equation as it did for the Galton-Watson process. However, it will be possible to estimate the random variable Y_{n+1} in the following sense. Suppose one wishes to choose a function $h\left(Y_n\right)$ such that the expectation $E\left[\left(Y_{n+1} - h\left(Y_n\right)\right)^2\right]$ is a minimum. It is known that $h\left(Y_n\right) = E\left[Y_{n+1} \mid Y_n\right]$ is the function that minimizes this expectation, and it is called the minimum $MEAN$ square estimate of the random variable Y_{n+1}. For $n = 1$ in (9.4.4), it can be seen that if $Y_1 = y_1$, then the estimate of Y_2 is

$$\widehat{Y}_2 = E\left[Y_2 \mid Y_1 = y_1\right] = y_1 \lambda p(y_1). \tag{9.4.5}$$

Thus, if we choose

$$\widehat{Y}_3 = \widehat{Y}_2 \lambda p(\widehat{Y}_2) \tag{9.4.6}$$

as an estimate of the random variable Y_3, we can continue this iteration to arrive at a nonlinear difference equation of the form

$$\widehat{Y}_{n+1} = \widehat{Y}_n \lambda p(\widehat{Y}_n) \tag{9.4.7}$$

for $n = 0, 1, 2, \ldots$ and $\widehat{Y}_0 = 1$. It should be noted that for $n > 2$, the estimates $\widehat{Y}_{n+1}$ may not be the minimum $MEAN$ square estimate of the random variable Y_{n+1}, but, nevertheless, it is of interest to iterate equation (9.4.7) and analyze its evolutionary behavior, using a mathematical analysis.

Equation (9.4.7) is a nonlinear difference equation and may be viewed as a deterministic model embedded in a stochastic process. As we will see, the analysis of this difference equation, will lead to considering iterates of a function of the form

$$f(y) = y\lambda p(y), \tag{9.4.8}$$

which is a mapping of the set of nonnegative real numbers $y \in \mathbf{R}^+ = [0, \infty)$ into itself. Just as in the . classical Galton-Watson process, where a linear function of the form $f(y) = y\lambda$ arises, the functional form in (9.4.8) will play an interesting role in both the mathematical and numerical analysis of a self-regulating branching process.

In a self-regulating population, it will be assumed that survivability of an individual depends on population size y. Given this assumption, it seems reasonable to assume that the greater the population size, the smaller is the probability an individual survives to reproduce. A plausible assumption, therefore, to place on the survival probability $p(y)$ is, therefore, that it is a monotone decreasing function of y, an assumption that leads one to consider survival functions. A survival function is any monotone decreasing function $s(y)$ with domain $[0, \infty)$ and range $[0, 1]$, satisfying the conditions $s(0) = 1$ and $s(y) \downarrow 0$ as $y \uparrow \infty$. Another factor to take into account when formulating a model of a self-regulating population is that survivability depends on the carrying capacity of the environment or on a conscientious desire to limit population size. Accordingly, one is led to consider a nonnegative parameter β and survival probabilities of the form $p(y) = s(\beta y)$, where β is related to the carrying capacity of the environment. Observe that if $\beta = 0$, then there is no density dependence and the process reduces to the classical Galton-Watson case. Furthermore, the smaller the value of β, the greater is the carrying capacity of the environment.

Survival functions form a very large class, making it necessary to focus on some explicit forms. A widely used parametric family of survival functions is the Weibull and in this case the survival probability takes the form

$$p_1(y) = \exp\left[-(\beta y)^\alpha\right], \tag{9.4.9}$$

where α is a positive shape parameter. For $\alpha \in [1, \infty)$, the function in (9.4.9) decreases more rapidly to 0 as $y \uparrow \infty$ for any β, than for the case $\alpha \in (0, 1)$. A question that naturally arises is whether a parametric family of another form should also be considered for the survival probability, giving rise to a source of uncertainty in formulating the model. One approach to dealing with this uncertainty is to consider another parametric family of survival functions and compare the implications of the two parametric families.

A very large class of survival functions consists of all Laplace transforms of continuous probability density functions $g(x)$ with support $x \in (0, \infty)$. In particular, if one considers the standard gamma density,

$$g(x) = \frac{1}{\Gamma(\gamma)} x^{\gamma-1} \exp\left[-x\right], y \in (0, \infty), \tag{9.4.10}$$

then the Laplace transform takes the form

$$s(y) = \int_0^\infty e^{-xy} g(x)dx = \frac{1}{(1+y)^\gamma}, \tag{9.4.11}$$

where $\Gamma(\gamma)$ is the gamma function, γ is a positive parameter, and $y \in [0, \infty)$ This Laplace transform gives rise to a family of survival probabilities of the form

$$p_2(y) = s\left(\beta y\right) = \frac{1}{(1+\beta y)^{\gamma}}, \tag{9.4.12}$$

such that for $\gamma \in (0,1)$, $p_2(y)$ converges rather slowly to 0 as $y \uparrow \infty$, but if $\gamma \in [1, \infty)$ convergence to 0 would be much more rapid. In the next section, both of these parametric families will be considered in developing theories for projecting total population size according to the embedded deterministic model in (9.4.7).

9.5 Fixed Points and Domains of Attraction for the Embedded Deterministic Model

If, given an initial number $y \in [0, \infty)$, the population is projected according to the embedded deterministic model, then projected population size in successive generations $n = 1, 2, 3, \ldots$ is determined by iterates of the function $f(y) = y\lambda p\left(y\right)$ in (9.4.7). Let $f_1(y) = f(y)$ and for $n \geq 2$ define $f_n(y)$ recursively by $f_n(y) = f(f_{n-1}(x))$. Then, projected population size in generation n is $f_n(y)$. If $f(\cdot)$ is a continuous function and these iterates converge to a limit y_ℓ, then this limit is a fixed point and a solution of the equation

$$y = f(y) = y\lambda p\left(y\right). \tag{9.5.1}$$

From (9.5.1) it can be seen that $y = 0$ is always a fixed point of the function, but there may also be other fixed points. It should be noted that none of the results of this section depend on the assumption the random variable N has a Poisson distribution, see (9.4.3). When the Weibull survival function in (9.4.9) is in force and $\beta > 0$, then it can be shown that a nonzero fixed point of the function is

$$y_f = \frac{(\ln \lambda)^{\frac{1}{\alpha}}}{\beta}. \tag{9.5.2}$$

Similarly, when the survival probability in (9.4.12) is in force and $\beta > 0$, then a nonzero fixed point of the function takes the form

$$y_f = \frac{\lambda^{\frac{1}{\gamma}} - 1}{\beta}. \tag{9.5.3}$$

For either of these fixed points, $y_f \in (0, \infty)$, if, and only if, $\lambda \in (1, \infty)$. From the biological point of view, the formulas in (9.5.2) and (9.5.3) may be interpreted as the carrying capacity of the environment; moreover, from these formulas, it is clear that the smaller the value of β, the greater is the carrying capacity. If the individuals of the population are consciously limiting the number of births per individual, then these formulas provide measures population size when the population is equilibrium.

A question that naturally arises is: for what points (α, β, λ) in the parameter space of the model and initial points $y \in (0, \infty)$ will the iterates of the function $f(y)$ converge to the fixed point in (9.5.2), when the Weibull survival function is in force? An analogous question may be asked for the parameter points (β, γ, λ) pertaining to (9.5.3). Let $f'(y)$ be the derivative of the function at the point y. According to widely used terminology, a fixed point y is said to be attracting if $\mid f'(y) \mid< 1$ and repelling if $\mid f'(y) \mid> 1$. For either survival function under consideration, it is easy to see that $f'(0) = \lambda$. Therefore, the fixed point $y = 0$ is either attracting or repelling according to whether $\lambda < 1$ or $\lambda > 1$. It is thus clear that the fixed points in (9.5.2) and (9.5.3) can be attracting only if $\lambda > 1$. By deriving a formula for $f'(y_f)$, the derivative of $f(y)$ evaluated at a fixed point y_f, points if the parameter space may be characterized as attracting. Such sets in the parameter spaces will be referred to as parametric domains of attraction. When the Weibull survival function is in force, then it can be shown that

$$f'(y_f) = 1 - \alpha \ln \lambda. \tag{9.5.4}$$

Therefore the condition $\mid f^{'}(y_f) \mid< 1$ is equivalent to

$$-1 < 1 - \alpha \ln \lambda < 1 \tag{9.5.5}$$

which in turn is equivalent to

$$1 < \lambda < \exp\left[\frac{2}{\alpha}\right]. \tag{9.5.6}$$

Therefore, for $\beta > 0$ the fixed point in (9.5.2) is attracting if, and only if, condition (9.5.6) is satisfied.

When the survival function in (9.4.12) is in force, it can be shown that

$$f^{'}(y_f) = \lambda^{-\frac{1}{\gamma}} + (1 - \lambda^{-\frac{1}{\gamma}})(\gamma - 1). \tag{9.5.7}$$

Therefore, the condition $\mid f'(y_f) \mid< 1$ is equivalent to

$$0 < \gamma(1 - \lambda^{-\frac{1}{\gamma}}) + 2\lambda^{-\frac{1}{\gamma}} < 2. \tag{9.5.8}$$

By using the condition $\gamma > 0$, it follows that this last inequality is equivalent to

$$0 < \gamma(1 - \lambda^{-\frac{1}{\gamma}}) < 2(1 - \lambda^{-\frac{1}{\gamma}}), \tag{9.5.9}$$

Then, by using the conditions $\gamma > 0$ and $\lambda > 1$, it can be seen that this inequality is equivalent to

$$0 < \gamma < 2. \tag{9.5.10}$$

Therefore, for $\beta > 0$ and $\lambda > 1$ the fixed point in (9.5.3) is attracting if, and only if, condition (9.5.11) is satisfied.

It is of interest to observe that the domain of attraction, as defined by condition (9.5.7) for the fixed point in (9.5.2) does not depend on the parameter β, and neither does the domain of attraction defined by (9.5.10) for fixed point (9.5.3). This observation is very useful in designing exploratory numerical experiments involving iterates of the function $f(y)$ in the two cases under study, for in these experiments it suffices to let $\beta = 1$. If, however, one is interested in these values when designing Monte Carlo simulation experiments using (9.5.2) and (9.5.3), then the value chosen for β would be essential in attempting to predict what the eventual size of a population may be when the number of generations in a projection is large.

For example, for formula (9.5.2) suppose $\alpha = 2$, then $e^{\frac{2}{\alpha}} = 2.718\,281\,828\,459\,05$, and from (9.5.7) it can be seen that if λ is assigned the value $\lambda = 2$, then the fixed point in (9.5.2) will be attracting for any $\beta > 0$. In particular, if $\beta = 10^{-6}$, then

$$\frac{(\ln \lambda)^{\frac{1}{\alpha}}}{\beta} = 832{,}554.611\,15 \tag{9.5.11}$$

so that, according to the embedded deterministic model, population size would converge to a limit of about 832, 555 individuals. On the other hand, if γ is chosen as a value in the interval $(0, 2)$, see (9.5.11), such as $\gamma = 1.5$, then the fixed point in (9.5.3) will be attracting and has the value

$$\frac{\lambda^{\frac{1}{\gamma}} - 1}{\beta} = 587{,}401.051\,96. \tag{9.5.12}$$

Interestingly, when the survival function defined in (9.4.11) is used in the model, then population as projected by the deterministic model would converge to a population of about 587, 401 individuals, which is smaller than that when the Weibull survival function was in force, see (9.5.11). At this point in the development of the mathematics underlying evolutionary theories of self-regulating branching process, it is not known, from a formal point of view, whether these fixed points are measures of central tendency for fluctuation of the sample functions of the stochastic process, but by conducting Monte Carlo simulation experiments and summarizing a sample of the sample functions statistically, it could be observed empirically whether these fixed points were indeed measures of central tendency of the stochastic process. Such experiments would also be helpful in designing Monte Carlo simulation experiments to simulate a sample of genealogies up to some generation n, starting with one initial individual. For if the fixed points are attracting, then by calculating their values, given some assignment of parameters in a parametric domain of attraction, it would be possible to estimate population size in some generation n so that some decisions could be made as to whether a simulated array $\mathcal{I}_n(\omega)$ of genealogies could be stored in a computer.

It thus appears that self-regulating branching process will have an advantage for simulating genealogies, when compared to Galton-Watson process in which population sizes grow without bounds when the expected number of offspring produced by each individual is greater than one. When fixed points are attracting and the iterates of the function governing the embedded deterministic model converge to the fixed point, then it could be said that the population is in equilibrium according to the embedded deterministic model. Just as with the discrete logistic model and others discussed by Gulick (1992), one would expect the iterates of both the embedded deterministic models to exhibit bifurcation, period doubling, and chaos at points in the parameter spaces outside the domains of attraction described in this section. In the computer implementation of the self-regulating branching process under consideration, some results of this kind will be reported in a subsequent section.

9.6 Probabilities of Extinction

All of the results of this section depend on the assumption the random variable N has a Poisson distribution. Unlike deterministic models, in stochastic formulations an embedded deterministic model may not completely characterize the behavior of the sample functions of the process. A case in point is the possibility a population becomes extinct. For $y \geq 1$, let

$$q(y,n) = P\left[Y_n = 0, Y_s \neq 0, 0 < s < n \mid Y_0 = y\right] \tag{9.6.1}$$

be the conditional probability that the population becomes extinct in generation $n \geq 1$, given that initial population size was y. When $n = 1$, this conditional probability has an explicit form for the two survival probabilities under consideration. For the case of the Weibull model, this probability takes the form

$$q_1(y,1) = \exp\left[-\lambda y \exp\left[-(\beta y)^{\alpha}\right]\right], \tag{9.6.2}$$

and in the other case, we have

$$q_2(y,1) = \exp\left[\frac{-\lambda y}{(1+\beta y)^{\gamma}}\right], \tag{9.6.3}$$

see (9.4.9) and (9.4.12). The conditional probability of extinction, given an initial population size $y \geq 1$, is

$$q(y) = \sum_{n=1}^{\infty} q(y,n). \tag{9.6.4}$$

A problem of interest in working with a self-regulating branching process is that of providing insights into values of $q(y)$. In principle, it follows from the properties of Markov chains with stationary transition probabilities that the probabilities in (9.6.1) may be calculated recursively. For if the absorbing state 0 is not entered at the first step, then some non-absorbing state $y \geq 1$ is entered on the first step and the process begins anew. Therefore, for $n \geq 2$ and $y \geq 1$ the probabilities in (9.6.1) satisfy the recursive system

$$q(y,n) = \sum_{z=1}^{\infty} p_{yx} q(z, n-1). \tag{9.6.5}$$

By summing these equations for $n \geq 2$ and using (9.6.4), it follows that for $y \geq 1$ the extinction probabilities satisfy

$$q(y) = q(y,1) + \sum_{z=1}^{\infty} p_{yz} q(z), \tag{9.6.6}$$

which is an infinite system of linear equations. Among others, Kemeny *et al.* (1966) have conducted theoretical studies of solutions to such infinite systems of linear equations. Although such studies are valuable, in practical applications it is useful to either calculate these probabilities, which may be very difficult, or to have useful non-zero lower bounds that may easily be calculated.

When the Weibull survival probability is in force, it can be seem from (9.6.2) that the existence of the limit,

$$\lim_{y \uparrow \infty} y p_1(y) = \lim_{y \uparrow \infty} y \exp[-(\beta y)^{\alpha}], \tag{9.6.7}$$

must be studied. Without loss of generality we may let $\beta = 1$. By passing to logarithms on the right, it can be seen that the existence of the limit depends on the behavior of

$$\ln(y) - y^{\alpha} = y^{\alpha}\left(\frac{\ln y}{y^{\alpha}} - 1\right) \tag{9.6.8}$$

as $y \uparrow \infty$. By L' Hospital's rule

$$\lim_{y\uparrow\infty}\frac{\ln y}{y^{\alpha}} = \lim_{y\uparrow\infty}\frac{1}{\alpha y^{\alpha}} = 0 \tag{9.6.9}$$

for every $\alpha > 0$. Therefore, because the right hand side of (9.6.8) becomes negative as $y \uparrow \infty$, it follows that

$$\lim_{y\uparrow\infty}(\ln y - y^{\alpha}) = -\infty \tag{9.6.10}$$

for every $\alpha > 0$, which implies that

$$\lim_{y\uparrow\infty} q_1(y,1) = \lim_{y\uparrow\infty}\exp(-\lambda y\exp -(\beta y)^{\alpha}) = 1. \tag{9.6.11}$$

Thus, when initial population size is large, it is virtually certain that the population will become extinct, because $1 = \lim_{y\uparrow\infty} q_1(y,1) \le q\,(y) \le 1$ so that $q\,(y) = 1$. This result is in sharp contract to the case when $\beta = 0$, so that the self-regulating process reduces to a Galton-Watson process. For in this case,

$$\lim_{y\uparrow\infty} q_1(y,1) = \lim_{y\uparrow\infty}\exp(-\lambda y) = 0, \tag{9.6.12}$$

indicating for that this model, if initial population size is large, the probability of extinction in the first generation is very small.

From (9.6.3), it can be seen that the behavior of $q_2\,(y,1)$ depends on the behavior of the limit

$$\lim_{y\uparrow\infty} yp_2(y) = \lim_{y\uparrow\infty}\left(\frac{y}{(1+\beta y)}\right)\left(\frac{1}{(1+\beta y)^{\gamma-1}}\right). \tag{9.6.13}$$

It can be seen, that if $\gamma > 1$, then this limit is 0 so that $q_2\,(y,1) \to 1$ as $y \uparrow \infty$. This is also a case in which it is virtually certain that the population will become extinct, given any large initial population size y. When $\gamma = 1$, however, this function approaches the limit $1/\beta$ as $y \uparrow \infty$. Therefore, for this case,

$$\lim_{y\uparrow\infty} q_2\,(y,1) = \exp\left(-\frac{\lambda}{\beta}\right). \tag{9.6.14}$$

Observe that when β is small, this probability is also small. Finally, if $\gamma \in (0,1)$, then the limit if (9.6.13) is infinite so that in this case $q_2\,(y,1) \to 0$ as $y \uparrow \infty$.

These results have interesting implications. For example, it is of interest to note that if for some choices of parameter values such that a positive fixed point is attracting and the initial population size is much larger than the carrying capacity of the environment, then, in a projection using the embedded deterministic model, population size will decrease drastically in the first generation but as time increases it would approach a fixed point, representing the carrying capacity of the environment. Yet, the integer valued random variable Y_n for the branching process will converge to 0 with probability one as $n \uparrow \infty$, indicating that the population actually becomes extinct. Thus, even though the embedded deterministic model and the branching process have the same parameters, in some cases they can behave quite differently.

9.7 Simulating Stochastic Genealogies in the Multitype Case with Mutation and Selection

If mutation occurs in some individual in a population, then, by definition, there is more than one type of individual in the population. Furthermore, if there are two or more types of individuals or gametes in a population, then selection may favor one of these types, and if, somehow, individuals of a given type on average contribute more offspring to the population in the next generation than those of other types, then this type would have a selective advantage over the others. In this section, the simulation of genealogies in multitype populations with mutation and selection will be considered with a view towards labeling individuals in a genealogy such that if any two individuals in some generation are selected at random, then it would be possible to find their most recent ancestor.

Suppose there are $k \geq 2$ types of individuals in a population and suppose a genealogy evolves from an initial individual of type i_0 denoted by $\left(i_0\ 0\right)$, where $i_0 = 1, 2, \ldots, k$. Just as with the case of one type of individual in a population, it will be instructive to consider some illustrative genealogies so as to provide insights into a more general case. Suppose for example, the initial individual $\left(i_0\ 0\right)$ contributes two offspring to generation 1. Then, the genealogy of the process in generation 1 may be represented as the 2×4 array

$$\mathcal{I}_1\left(\omega_1\right) = \begin{pmatrix} i_0 & 0 & i_1 & 1 \\ i_0 & 0 & i_2 & 2 \end{pmatrix}, \tag{9.7.1}$$

where, for example, if $i = i_1$, then no mutation occurred in the initial individual, but if $i \neq i_1$, then a mutation did occur. The same statement holds of the second row or individual of the array. As indicated in the array, the individual $\left(i_0\ 0\ i_1\ 1\right)$ in the first generation is an offspring of individual $\left(i_0\ 0\right)$ in the initial generation, and the second row of the array is interpreted similarly.

Next suppose that each individual of generation 1 also contributes two offspring to generation 2. Then, the genealogy of the process in generation 2 may be represented by 4 × 6 array

$$\mathcal{I}_2\left(\omega_2\right) = \begin{pmatrix} i_0 & 0 & i_1 & 1 & i_3 & 3 \\ i_0 & 0 & i_1 & 1 & i_4 & 4 \\ i_0 & 0 & i_2 & 2 & i_5 & 5 \\ i_0 & 0 & i_2 & 2 & i_6 & 6 \end{pmatrix}. \tag{9.7.2}$$

In this array, the individual $\left(i_0\ 0\ i_1\ 1\ i_3\ 3\right)$ in generation 2 may be viewed as the first offspring of individual $\left(i_0\ 0\ i_1\ 1\right)$ in generation 1. The other three rows in the array have a similar interpretations.

From the array $\mathcal{I}_2\left(\omega_2\right)$, it can be seen that the most recent ancestor for individuals $\left(i_0\ 0\ i_1\ 1\ i_3\ 3\right)$ and $\left(i_0\ 0\ i_1\ 1\ i_4\ 4\right)$ is the individual $\left(i_0\ 0\ i_1\ 1\right)$ in generation 1. Similarly, the most recent common ancestor of the individuals $\left(i_0\ 0\ i_1\ 1\ i_3\ 3\right)$ and $\left(i_0\ 0\ i_2\ 2\ i_6\ 6\right)$ is the initial individual $\left(i_0\ 0\right)$. Let $\left(i_0\ 0\ i_{\nu_1}\ \nu_1\ i_{\nu_2}\ \nu_2\right)$ denote an arbitrary individual in the array $\mathcal{I}_2\left(\omega_2\right)$. Then, the condition $i_0 = i_{\nu_1} = i_{\nu_2}$ indicates that no mutation has occurred in this line of descent from the initial individual $\left(i_0\ 0\right)$. If this condition does not hold, then there is at least one mutation in this line of descent. To help fix ideas, the last pair of indices in the individual $\left(i_0\ 0\ i_{\nu_1}\ \nu_1\ i_{\nu_2}\ \nu_2\right)$ will be of type i_{ν_2}. From the point of view of genetics, the idea of type could also be interpreted as haplotype.

It is also clear that the schemes presented in (9.7.1) and (9.7.2) could also be generalized to accommodate some total number $y_n \geq 1$ of individuals in generation $n \geq 1$. Let $\mathcal{I}_n\left(\omega_n\right)$ denote an array of y_n individuals in generation n. Then, an arbitrary individual ι in this array may be represented as the vector of pairs of numbers

$$\iota = \left(i_0, 0, i_{\nu_1}, \nu_1, \ldots, i_{\nu_n}, \nu_n\right).$$

Altogether, this vector contains $2\left(n+1\right)$ numbers describing a line of descent from the initial individual to individual ι in generation n. The vector $\boldsymbol{i} = \left(i_0, i_{\nu_1}, \ldots, i_{\nu_n}\right)$ indicates the types of individuals in this line of descent, where $i_{\nu_\xi} = 1, 2, \ldots, k$ for $\xi = 1, 2, \ldots, n$, and the vector $\left(0, \nu_1, \ldots, \nu_n\right)$ is a set of numbers that identifies the line of descent of individual ι uniquely in the array $\mathcal{I}_n\left(\omega_n\right)$ for generation n. To further illustrate these ideas, let ι and ι' denote two individuals in the array $\mathcal{I}_n(\omega_n)$. Then if $\nu_k = \nu'_k$ for $k = 0, 1, \ldots, n_0 < n$ but $\nu_k \neq \nu'_k$ for $k = n_0 + 1, \ldots, n - 1$, then the individual $(\nu_0, \nu_1, \ldots, \nu_{n_0})$ in generation n_0 is the most recent ancestor of ι and ι'. Moreover, if $i_{n_0} = i'_{n_0}$, then this ancestor was of type i_{n_0}, but it may also be the case that $i_{n_0} \neq i'_{n_0}$so that the types of individuals in the lines of descent of individuals ι and ι' up to generation n_0 may differ. Thus, with this scheme of labeling individuals in a genealogy or sets of lines of descent, the most recent ancestor of two individuals may be unambiguously defined but the types of individuals among lines of decent may vary when two individual in some generation n are compared. If, however, for some individual ι in generation n the vector $\boldsymbol{i}$ has the property $i_0 = i_{\nu_k}$ for $k = 1, 2, \ldots, n$, then there has been no mutations in the line of descent leading up to individual ι in generation n. When simulating a set of lines of descent in a sample of genealogies, mutations are rare events. Therefore, this observation could form the basis for writing software to screen for individuals in the array $\mathcal{I}_n(\omega_n)$ for mutations in their lines of descent. In such a sample of genealogies, it would not be surprising if many lines of descent contained no mutations.

A class of multitype branching processes that may be used to simulate a sample of multitype genealogies is often referred to as a multitype Galton-Watson process, which is defined as follows. Suppose a process has $k \geq 2$ types and let $\boldsymbol{X}_\nu = (X_{\nu 1}, X_{\nu 2}, \ldots, X_{\nu k})$ for $\nu = 1, 2, \ldots, k$ denote a collection of independent $1 \times k$ vector valued variables taking values in the set

$$\mathfrak{S}_k = ((x_1, x_2, \ldots, x_k) \mid x_\nu = 0, 1, 2, \ldots \text{ for } \nu = 1, 2, \ldots, k) \qquad (9.7.3)$$

of k-dimensional vectors of nonnegative integers. Each of these random vectors has the following interpretation. If there is some individual of type ν in some generation n, then the numbers of offspring contributed by this individual to generation $n + 1$ of each of the k types is a realization of the random vector $\boldsymbol{X}_\nu$ taking values in the set $\mathfrak{S}_k$. As will be illustrated subsequently, this set may also be interpreted as the state space of a vector valued Markov chain, which may be defined as follows.

Let $\boldsymbol{Y}(n) = (Y_1(n), Y_2(n), \ldots, Y_k(n))$ for $n = 0, 1, \ldots$ be sequence of vector valued random variable taking values in the set $\mathfrak{S}_k$, denoting the numbers of individuals of each type present in a population in generation n. Suppose that for the initial population $\boldsymbol{Y}(0) = \boldsymbol{y}(0) \in \mathfrak{S}_k$, an arbitrary vector $1 \times k$ vector of nonnegative integers. Just as with the one type Galton-Watson process, realizations of these vector valued random variables may be computed recursively. For a given vector

$\boldsymbol{Y}(n-1) = (Y_1(n-1), Y_2(n-1), \ldots, Y_k(n-1))$ in generation $n-1$, let $(\boldsymbol{X}_{\nu j})$ for $j = 1, 2, \ldots, Y_\nu(n-1)$, be a collection of conditionally independent vector valued random variables with the same distribution as the vector $\boldsymbol{X}_\nu$ for $\nu = 1, 2, \ldots, k$. Then, the vector $\boldsymbol{Y}(n)$ in generation n is given by the random sums

$$\boldsymbol{Y}(n) = \sum_{\nu=1}^{k} \sum_{j=1}^{Y_\nu(n-1)} \boldsymbol{X}_{\nu j}, \tag{9.7.4}$$

where if $Y_\nu(n-1) = 0$ for some ν, then the corresponding sum in (9.7.4) is interpreted as the zero vector. In the next section, a more detailed theoretical structure of these random sums will be outlined. When simulating a genealogy in this multitype case, the values of each of the vector random variables $\boldsymbol{X}_{\nu j}$ would need to be included in the software, because they represent the offspring of each individual.

9.8 On Parameterizing Multitype Galton-Watson Processes

In order to develop a theoretical version of a multitype Galton-Watson process described in the previous section in a form that can be implemented on computers, it will be helpful to explore a way in which the model may be parameterized to accommodate mutation and selection. For every $\boldsymbol{x} \in \mathfrak{S}_k$, let

$$P[\boldsymbol{X}_\nu = \boldsymbol{x}] = p_\nu(\boldsymbol{x}) \tag{9.8.1}$$

for $\nu = 1, 2, \ldots, k$. Then, the probability generating function of the random vector $\boldsymbol{X}_\nu$ is defined for every real or complex vector $\mathbf{s} = (s_1, s_2, \ldots, s_k)$ such that $\max_\nu \mid s_\nu \mid \leq 1$ by the expression

$$f_\nu(\mathbf{s}) = \sum_{\boldsymbol{x}} p_\nu(\boldsymbol{x}) s_1^{x_1} s_2^{x_2} \ldots {}_k^{x_k}, \tag{9.8.2}$$

where the sum extends over all $\boldsymbol{x} \in \mathfrak{S}_k$ for $\nu = 1, 2, \ldots, k$. In the remainder of this section, some schemes for parameterizing these generating functions will be given in terms of mutation and selection.

Just as with the formalization of multitype gametes sampling models presented in chapter 5, mutation among types will be accommodated in the model by introducing the $k \times k$ matrix of mutation probabilities

$$\mathfrak{M} = (\mu_{\nu\nu'}), \tag{9.8.3}$$

where $\mu_{\nu\nu'} \geq 0$ for all pairs (ν, ν') and

$$\sum_{\nu'=1}^{k} \mu_{\nu\nu'} = 1. \tag{9.8.4}$$

for every $\nu = 1, 2, \ldots, k$. Note the matrix $\mathfrak{M}$ could also be interpreted as a transition matrix for a Markov chain.

Let the random variable N_ν taking values in the set of nonnegative integers denote the total number of offspring produced by an individual of type ν in any generation. Then, let δ_{ij} denote the Kronecker delta defined by $\delta_{ii} = 1$ and $\delta_{ij} = 0$ if $i \neq j$, and let $\boldsymbol{\varepsilon}_\nu = (\delta_{1\nu}, \delta_{2\nu}, \ldots, \delta_{k\nu})$ for $\nu = 1, 2, \ldots, k$ denote a set of basis vectors for a Euclidean space of k dimensions. Next, let $\boldsymbol{\xi}_\nu$denote a generalized Bernoulli indicator random variable indicating taking values in the set $(\boldsymbol{\varepsilon}_\nu \mid \nu = 1, 2, \ldots, k)$. Observe that

$$P\left[\boldsymbol{\xi}_\nu = \boldsymbol{\varepsilon}_{\nu'}\right] = \mu_{\nu\nu'} \tag{9.8.5}$$

and the generating function of $\boldsymbol{\xi}_\nu$ is

$$g_\nu\left(s_1, s_2, \ldots, s_k\right) = \sum_{\nu'=1}^{k} \mu_{\nu\nu'} s_{\nu'}. \tag{9.8.6}$$

For each $\nu = 1, 2, \ldots .k$, let $\left(\boldsymbol{\xi}_{\nu j} \mid j = 1, 2, \ldots, N_v\right)$ denote a collection of conditionally independent random variables whose common distribution is that of $\boldsymbol{\xi}_\nu$, given N_ν. Then, the vector random variable $\boldsymbol{X}_\nu$, representing the number of offspring of each type produced by an individual of type ν, may be represented as the vector valued random sum

$$\boldsymbol{X}_\nu = \sum_{j=1}^{N_\nu} \boldsymbol{\xi}_{\nu j}. \tag{9.8.7}$$

Selection in this formulation will be accommodated in terms of the finite expectations $\lambda_\nu = E\left[N_\nu\right]$ of the random variables in $(N_1, N_2, \ldots, N_k)$. Let $h_\nu(s)$ denote the generating function

$$h_\nu(s) = \sum_{n=0}^{N_\nu} P\left[N_\nu = n\right] s^n \tag{9.8.8}$$

of the random variable N_ν for $\mid s \mid \leq 1$. Then, from (9.8.6) and (9.8.7), it follows that

$$f_\nu(\mathbf{s}) = h_\nu\left(g_\nu\left(s_1, s_2, \ldots, s_k\right)\right) \tag{9.8.9}$$

for $\nu = 1, 2, \ldots, k$. In particular, if, for every ν, the generating function $h_\nu(s)$ has the Poisson form, then

$$h_\nu(s) = \exp(\lambda_\nu(s-1)), \tag{9.8.10}$$

and

$$f_\nu(\mathbf{s}) = \exp\left(\lambda_\nu \sum_{\nu'=1}^{k} \mu_{\nu\nu'}(s_{\nu'}-1)\right), \tag{9.8.11}$$

where $\lambda_\nu > 0$ for $\nu = 1, 2, \ldots, k$. This parametric form of the offspring distributions for a multitype branching process is particularly useful in computer implementations of the model.

For example, let $\boldsymbol{x}_\nu$ denote a realization of the vector random variable $\boldsymbol{X}_\nu$ representing the number of offspring of each type produced by an individual of type ν is any generation. From the random sum in (9.8.7), it can be seen that to compute the realization $\boldsymbol{x}_\nu$, the first step would be that of computing a realization n_ν of the random variable N_ν. Then, given that $N_\nu = n_\nu$, the vector $\boldsymbol{x}_\nu$ could be computed as a sample from a multinomial distribution with index n_ν and probability vector $(\mu_{\nu 1}, \mu_{\nu 2}, \ldots, \mu_{\nu k})$. When simulating a genealogy based on the evolution of a multitype branching process, this step would have to carried out for each individual in some generation $n-1$. To illustrate this remark, suppose, for example, in generation $n-1$ that $\boldsymbol{y}(n-1) = (y_1(n-1), y_2(n-1), \ldots, y_k(n-1))$ is the realization of the random vector $\boldsymbol{Y}(n-1)$ so that total population size in this generation is $y(n-1) = y_1(n-1) + y_2(n-1) + \cdots + y_k(n-1)$. Then, in generation $n-1$, the array $\mathcal{I}_{n-1}(\omega_{n-1})$ would have dimensions $y(n-1) \times 2n$. Let ι denote the individual in row 1 of the array $\mathcal{I}_{n-1}(\omega_{n-1})$ and suppose this individual is of type ν. Then, a realization $\boldsymbol{x}_\nu = (x_{\nu 1}, x_{\nu 2}, \ldots, x_{\nu k})$ of the random vector $\boldsymbol{X}_\nu$ would have to be computed. Given this realization, $x_{\nu 1}$ is the number of offspring of type 1 that individual ι in generation $n-1$ contributes to generation n, $x_{\nu 2}$ is the number of offspring of type 2 contributed to the next generation and so on down to $x_{\nu k}$, the number of individuals of type k contributed to the next generation. This process would continue until all the $y(n-1)$ individuals in the array $\mathcal{I}_{n-1}(\omega_{n-1})$ for generation $n-1$ had been included to produce the array $\mathcal{I}_n(\omega_n)$ for generation n. If $\boldsymbol{y}(n) = (y_1(n), y_2(n), \ldots, y_k(n))$ is the realized population vector for generation n so that total population size in this generation is $y(n) = y_1(n) + y_2(n) + \cdots + y_k(n)$, then the array $\mathcal{I}_n(\omega_n)$ would have dimensions $y(n) \times 2(n+1)$. The labeling of the individuals in the rows of array $\mathcal{I}_n(\omega_n)$ would be extensions of that described in section (9.7). If a decision were made not to simulate a genealogy, then the simulation of realizations of the $1 \times k$ vectors $\boldsymbol{Y}(n)$ for the population is a straight forward task. Suppose the realized population vector in generation $n-1$ is $\boldsymbol{y}(n-1) = (y_1(n-1), y_2(n-1), \ldots, y_k(n-1))$, and let the $1 \times k$ vector $\boldsymbol{x}_\nu$ denote a sample of size 1 from a multinomial distribution with index $y_\nu(n-1)$ and probability vector $(\mu_{\nu 1}, \mu_{\nu 2}, \ldots, \mu_{\nu k})$ for $\nu = 1, 2, \ldots, k$. Then, the realized population vector for generation n would be given by the vector sum

$$\boldsymbol{y}_n = \sum_{\nu=1}^{k} \boldsymbol{x}_\nu. \tag{9.8.12}$$

Observe that this equation is another way of writing equation (9.7.4). A $k \times k$ expectation matrix $\boldsymbol{\Lambda} = (\lambda_{\nu\nu'})$ plays a basic role the classification and analysis of multitype branching processes. The elements of this matrix are defined as follows. Let $\boldsymbol{X}_\nu = (X_{\nu 1}, X_{\nu 2}, \ldots, X_{\nu k})$ denote the $1 \times k$ vector sum defined in (9.8.7). Then, by definition, the expectation of this vector is

$$E[\boldsymbol{X}_\nu] = (E[X_{\nu 1}], E[X_{\nu 2}], \ldots, E[X_{\nu k}]), \tag{9.8.13}$$

and the expectation $\lambda_{\nu\nu'}$ is defined by $\lambda_{\nu\nu'} = E[X_{\nu\nu'}]$. From (9.8.7) it can be seen that

$$E[\boldsymbol{X}_\nu \mid N_\nu] = N_\nu (\mu_{\nu 1}, \mu_{\nu 2}, \ldots \mu_{\nu k}), \tag{9.8.14}$$

because $E[\boldsymbol{\xi}_{\nu l}] = (\mu_{\nu 1}, \mu_{\nu 2}, \ldots \mu_{\nu k})$ for every $j = 1, 2, \ldots, N_\nu$. Let $\lambda_\nu = E[N_\nu]$ denote the expectation of the random variable N_ν. Then, the unconditional expectation of the random vector $\boldsymbol{X}_\nu$ is

$$E[\boldsymbol{X}_\nu] = \lambda_\nu (\mu_{\nu 1}, \mu_{\nu 2}, \ldots \mu_{\nu k}) \tag{9.8.15}$$

so that the $k \times k$ matrix $\boldsymbol{\Lambda}$ has the form

$$\boldsymbol{\Lambda} = (\lambda_\nu \mu_{\nu\nu'}). \tag{9.8.16}$$

The elements of this matrix could also have been derived by differentiating the generating functions in (9.8.9) and (9.8.11). Observe all the elements of the matrix $\boldsymbol{\Lambda}$ are nonnegative so that the Perron-Frobenius theory of matrices with nonnegative elements applies.

The matrix $\boldsymbol{\Lambda}$ plays a basic role in finding the expectation of a random population vector $\boldsymbol{Y}(n)$ in any generation $n = 1, 2, \ldots$, given a fixed initial vector $\boldsymbol{y}(0)$. Let $\boldsymbol{\nu}(n) = E[\boldsymbol{Y}(n)]$ for $n = 1, 2, \ldots$. Then, by taking expectations in (9.7.4), it can be seen that

$$\boldsymbol{\nu}(n) = \boldsymbol{\nu}(n-1)\boldsymbol{\Lambda}, \tag{9.8.17}$$

and by iterating this equation, it can be seen that for any generation $n \geq 1$

$$\boldsymbol{\nu}(n) = \boldsymbol{y}(0)\,\boldsymbol{\Lambda}^{n}. \tag{9.8.18}$$

Thus, the behavior of the expectation vector is determined by the properties of the expectation matrix $\boldsymbol{\Lambda}$.

The properties of this matrix depend on its form of which there are many. One form of particular interest is the case in which the matrix $\boldsymbol{\Lambda}$ is such that for some integer $m \geq 1$, all the elements of $\boldsymbol{\Lambda}^{m}$ are positive. Such a matrix is said to be positively regular. For the case of matrices with the form in (9.8.16), if the mutation matrix $\mathfrak{M}$ is positively regular, then so is the matrix $\boldsymbol{\Lambda}$. When the matrix $\boldsymbol{\Lambda}$ is positively regular, then population growth can be characterized in terms of the eigenvalue $\rho > 0$, the Perron-Frobenius root of the matrix $\boldsymbol{\Lambda}$. In particular, if $\rho > 1$, then the total expected size of the population will increase without bound. But, if $\rho < 1$, then the total expected size of the population will decrease as n, the number of generations, increase.

Another case of interest is that when there are three types of individuals and the mutation matrix has the form

$$\mathfrak{M} = \begin{pmatrix} \mu_{11} & \mu_{12} & \mu_{13} \\ 0 & \mu_{22} & \mu_{23} \\ 0 & 0 & 1 \end{pmatrix}, \tag{9.8.19}$$

so that the expectation matrix has the form

$$\boldsymbol{\Lambda} = \begin{pmatrix} \lambda_1\mu_{11} & \lambda_1\mu_{12} & \lambda_1\mu_{13} \\ 0 & \lambda_2\mu_{22} & \lambda_2\mu_{23} \\ 0 & 0 & \lambda_3 \end{pmatrix}. \tag{9.8.20}$$

In this case, the matrix $\boldsymbol{\Lambda}$ is called reducible, and, as can be seen, it is not positively regular.

In this book no general theories of multitype branching processes will be considered, but if a reader is interested in such theories the books Harris (1963) and Mode (1971) may be consulted. In chapter 1 of the latter reference, an account for limit theorems for the case the matrix $\boldsymbol{\Lambda}$ is positive regular may be found in the subcases $\rho < 1, \rho = 1$ and $\rho > 1$. Information on extinction probabilities is also available in chapter 1 of this book. Moreover, in chapter 2 of this book, cases when the matrix $\boldsymbol{\Lambda}$ is reducible are also considered. That a multitype Galton-Watson process may be formulated as a Markov chain is also discussed in the latter book.

It is of interest to note that the eigenvalues of the matrix in (9.8.20) are the diagonal elements $\lambda_1\mu_{11}, \lambda_2\mu_{22}$ and λ_3, and if type 3 has a selective advantage over the other types so that $\lambda_3 > \lambda_2 > \lambda_1$, individuals of type 3 would eventually become fixed in the population. The growth rate for this population at fixation would be the parameter λ_3. In formulation under consideration, the rate of approach to fixation could also be obtained by considering the spectral decomposition of the matrix $\mathbf{\Lambda}$. For let $\boldsymbol{D} = diag\left(\lambda_1\mu_{11}, \lambda_2\mu_{22}, \lambda_3\right)$ of eigenvalues and let $\boldsymbol{T}$ denote a nonsingular matrix such that

$$\mathbf{\Lambda} = \boldsymbol{T}^{-1}\boldsymbol{D}\boldsymbol{T}. \tag{9.8.21}$$

Then, it follows that

$$\boldsymbol{\nu}\left(n\right) = \boldsymbol{y}\left(0\right)\boldsymbol{T}^{-1}\boldsymbol{D}^{n}\boldsymbol{T} \tag{9.8.22}$$

for all $n \geq 1$. Hence, the eigenvalues on the diagonal of the matrix $\boldsymbol{D}$ would determine the rate of growth of the expectation vector $\boldsymbol{\nu}\left(n\right)$ as n increases. If $\lambda_3 > 1$, then this vector would increase without bound. Therefore, from the perspective of simulating genealogies, a theory of a self-regulating multitype branching process would not only be of biological interest but it would also be of practical interest to avoid large arrays that could not be stored in a desk top computer. In the next section, such a process will be considered.

9.9 Self-Regulating Multitype Branching Processes

The formulation of the class of self-regulating branching processes to be considered in this section is an extension of the ideas used in the formulation self-regulating process for the case of only one type in section 9.2. Let $\boldsymbol{y}\left(n\right) = \left(y_1\left(n\right), y_2\left(n\right), \ldots, y_k\left(n\right)\right)$ denote a realization of the population vector $\boldsymbol{Y}\left(n\right)$ for a multitype branching process with $k \geq 2$ types in generation n. For $y \in \mathbb{R}_+ = [0, \infty)$, the set of nonnegative real numbers, let $s_\nu\left(y\right)$ denote some parametric survival function and let $p_\nu\left(\boldsymbol{y}\left(n\right)\right)$ denote the probability an individual of type ν in generation n survives to contribute offspring to generation $n+1$. Then, this survival probability will be chosen as

$$p_\nu\left(\boldsymbol{y}\left(n\right)\right) = s_\nu\left(\beta_\nu\gamma_\nu\left(\boldsymbol{y}\left(n\right)\right)\right), \tag{9.9.1}$$

where β_ν is a parameter such that $\beta_\nu \geq 0$ and $\gamma_\nu\left(\boldsymbol{y}\left(n\right)\right)$ is a function that maps a vector $\boldsymbol{y}\left(n\right)$ of nonnegative integers into the set $\mathbb{R}_+$ of nonnegative real numbers for each type $\nu = 1, 2, \ldots, k$. One simple choice for the function $\gamma_\nu\left(\boldsymbol{y}\left(n\right)\right)$ is $\gamma_\nu\left(\boldsymbol{y}\left(n\right)\right) = y_\nu\left(n\right)$ for $\nu = 1, 2, \ldots, k$, and another choice for any ν could be

$$\gamma_\nu\left(\boldsymbol{y}\left(n\right)\right) = y\left(n\right) = \sum_{\nu=1}^{k} y_\nu\left(n\right), \tag{9.9.2}$$

the total size of the population in generation n. To illustrate these ideas, suppose a Weibull survival function of the form

$$s_\nu\left(y\right) = \exp\left(-y^{\alpha_\nu}\right), \tag{9.9.3}$$

where $\alpha_\nu > 0$, is chosen for each type $\nu = 1, 2, \ldots, k$. Then, if the functions γ_ν are chosen as $\gamma_\nu\left(\boldsymbol{y}\left(n\right)\right) = y_\nu\left(n\right)$ for $\nu = 1, 2, \ldots, k$, the survival probabilities take the form

$$p_\nu\left(\boldsymbol{y}\left(n\right)\right) = \exp\left(-\left(\beta_\nu y_\nu\left(n\right)\right)^{\alpha_\nu}\right) \tag{9.9.4}$$

for each type ν. It is of interest to observe, that for this choice of the function γ_k, the survival probability of individual of type ν in generation n depends only on the number of individuals $y_\nu\left(n\right)$ of individuals of that type in the population in generation n. As will be seen subsequently, when all the survival probabilities have this form, explicit formulas for fixed point on an embedded deterministic model may be derived. On the other hand, if the function in (9.9.2) is chosen, then

$$p_\nu\left(\boldsymbol{y}\left(n\right)\right) = \exp\left(-\left(\beta_\nu y\left(n\right)\right)^{\alpha_\nu}\right) \tag{9.9.5}$$

is the survival probability for each type ν. Thus, for this choice of the function $\gamma_\nu\left(\boldsymbol{y}\left(n\right)\right)$, the survival probability is a function of total population size for each type $\nu = 1, 2, \ldots, k$. This choice for survival probabilities may be more realistic than those for the first choice of the function γ_k, because total population size seems to be a more plausible measure of the competition among individuals in a population as they compete for available resources.

Now suppose a multitype self-regulating branching process is formulated in a manner similar to that described in section 9.4, and for some generation $n \geq 1$, let

$$\boldsymbol{X}_\nu\left(n\right) = \left(X_{\nu 1}\left(n\right), X_{\nu 2}\left(n\right), \ldots, X_{\nu k}\left(n\right)\right) \tag{9.9.6}$$

denote a $1 \times k$ vector representing the numbers of offspring of each type contributed to generation $n + 1$ by an individual of type $\nu = 1, 2, \ldots, k$ in generation n. It will also be assumed that parameterization of the distribution of the random vectors follows the scheme presented in section 9.8 except that the expectation of a random variable $N_\nu\left(n\right)$, representing the total number of offspring an individual of type ν in generation n contributes to generation $n+1$, has the form

$$E\left[N_{\nu}\left(n\right)\right]=\lambda_{\nu}p_{\nu}\left(\boldsymbol{y}\left(n\right)\right), \tag{9.9.7}$$

see (9.9.1), where as in section 9.8, $\lambda_{\nu}>0$ is a parameter for $\nu=1,2,\ldots,k$. Then, if it is assumed that the mutation matrix defined in (9.8.3) is in force, it follows that the conditional expectation of the vector $\boldsymbol{X}_{\nu}\left(n\right)$, given that $\boldsymbol{Y}\left(n\right)=\boldsymbol{y}\left(n\right)$, is

$$E\left[\boldsymbol{X}_{\nu}\left(n\right)\mid\boldsymbol{Y}\left(n\right)=\boldsymbol{y}\left(n\right)\right]=\lambda_{\nu}p_{\nu}\left(\boldsymbol{y}\left(n\right)\right)\left(\mu_{\nu1},\mu_{\nu2},\ldots,\mu_{\nu k}\right) \tag{9.9.8}$$

for $\nu=1,2,\ldots,k$. Furthermore, from this equation, it can be seen that $k\times k$ conditional expectation matrix analogous to that in (9.8.16) takes the form

$$\boldsymbol{\Lambda}\left(\boldsymbol{y}\left(n\right)\right)=\left(\lambda_{\nu}p_{\nu}(\boldsymbol{y}\left(n\right))\mu_{\nu\nu'}\right) \tag{9.9.9}$$

for the class of self-regulating multitype branching process under consideration.

For $n=0,1,2,\ldots$ let $\left(\boldsymbol{Y}\left(n\right)\right)$ denote a sequence of population vectors describing the evolution of a multitype self-regulating branching process, given the initial vector $\boldsymbol{Y}\left(0\right)=\boldsymbol{y}\left(0\right)$. This sequence is a Markov chain with state space $\mathfrak{S}_{k}$, the set of k-dimensional vectors of nonnegative integers. The transition matrix of this Markov chain is defined in a manner similar to that in presented in section 9.4. Monte Carlo realizations of these random vectors may be computed following the procedures outlined in section 9.8 except that for each generation n, the survival probabilities $p_{\nu}\left(\boldsymbol{y}\left(n\right)\right)$ must be computed for each type $\nu=1,2,\ldots,k$. In generation $n+1$, the conditional expectation of the random vector $\boldsymbol{Y}\left(n+1\right)$, given that $\boldsymbol{Y}\left(n\right)=\boldsymbol{y}\left(n\right)$, is

$$E\left[\boldsymbol{Y}\left(n+1\right)\mid\boldsymbol{Y}\left(n\right)=\boldsymbol{y}\left(n\right)\right]=\boldsymbol{y}\left(n\right)\boldsymbol{\Lambda}\left(\boldsymbol{y}\left(n\right)\right) \tag{9.9.10}$$

for $n=0,1,2,\ldots$. In particular, for $n=0$, the right hand side of this equation can be calculated and it may be concluded that

$$\widehat{\boldsymbol{Y}}\left(1\right)=\boldsymbol{y}\left(0\right)\boldsymbol{\Lambda}\left(\boldsymbol{y}\left(0\right)\right) \tag{9.9.11}$$

is the best estimator of the random vector $\boldsymbol{Y}\left(n+1\right)$ is the sense of minimum $MEAN$ square error and can be easily calculated because the right hand side of (9.9.10) is known for $n=0$. Thus, just as with the case of a self regulatory branching process with one type, the sequence of random vectors $\left(\boldsymbol{Y}\left(n\right)\right)$ may be estimated recursively by using the equation for $n\geq1$

$$\widehat{\boldsymbol{Y}}\left(n+1\right)=\widehat{\boldsymbol{Y}}\left(n\right)\boldsymbol{\Lambda}\left(\widehat{\boldsymbol{Y}}\left(n\right)\right), \tag{9.9.12}$$

and letting $\widehat{\boldsymbol{Y}}\left(1\right)$ equal the expression on the right in (9.9.11).

This nonlinear difference equation, which is embedded in a multitype self-regulating branching process, may be viewed as a mapping of the set

$$\mathbb{R}_{+}^{k} = ((y_1, y_2, \ldots, y_k) \mid y_\nu \in \mathbb{R}_{+} \text{ for } \nu = 1, 2, \ldots, k) \tag{9.9.13}$$

of k-dimensional vectors on nonnegative real numbers into itself. To simplify the notation, for every $\boldsymbol{y} \in \mathbb{R}_{+k}$, define the mapping

$$\boldsymbol{T}(\boldsymbol{y}) = \boldsymbol{y}\boldsymbol{\Lambda}(\boldsymbol{y}) \tag{9.9.14}$$

of the set $\mathbb{R}_{+}^{k}$ into itself. Then, solving equation (9.9.13) numerically is equivalent for computing a sequence of vectors $\boldsymbol{y}(n), n = 0, 1, \ldots$, recursively using the equation

$$\boldsymbol{y}(n+1) = \boldsymbol{T}(\boldsymbol{y}(n)) \tag{9.9.15}$$

for $n \geq 0$, where $\boldsymbol{y}(0)$ is an assignment for the initial vector. If this sequence converges to a limit $\boldsymbol{y} \in \mathbb{R}_{+k}$, then, because $\boldsymbol{T}(\boldsymbol{y})$ is a continuous function for the class of models under consideration, this limit is a fixed point in the sense that

$$\boldsymbol{y} = \boldsymbol{T}(\boldsymbol{y}) = \boldsymbol{y}\boldsymbol{\Lambda}(\boldsymbol{y}). \tag{9.9.16}$$

Although it is frequently possible to verify numerically that a sequence $\boldsymbol{y}(n), n = 0, 1, \ldots$, converges to some vector $\boldsymbol{y} \in \mathbb{R}_{+}^{k}$, in general it will be possible to derive symbolic forms of this limit only in special cases. One such special case arises when it is assumed that there is no mutation so that the mutation matrix has the form $\mathfrak{M} = \boldsymbol{I}_k$, a $k \times k$ identity matrix. If it is also assumed that the functions $\gamma_\nu(\boldsymbol{y})$ have the form $\gamma_\nu(\boldsymbol{y}) = y_\nu$ for $\nu = 1, 2, \ldots, k$ so that the matrix $\boldsymbol{\Lambda}(\boldsymbol{y})$ has the diagonal form

$$\boldsymbol{\Lambda}(\boldsymbol{y}) = diag(\lambda_\nu p_\nu(y_\nu) \mid \nu = 1, 2, \ldots, k), \tag{9.9.17}$$

then equation (9.9.16) reduces to

$$y_\nu = y_\nu \lambda_\nu p_\nu(y_\nu) \tag{9.9.18}$$

for $\nu = 1, 2, \ldots, k$. From this equation, it can be seen that $y_\nu = 0$ for all $\nu = 1, 2, \ldots, k$ is a fixed point of this equation, and, indeed, as can be seen from the general equation in (9.9.16), the zero vector $\boldsymbol{y} = \boldsymbol{0}$ is always a fixed point of this equation. Furthermore, if $y_\nu > 0$ for all $\nu = 1, 2, \ldots, k$, then in this case a fixed vector $\boldsymbol{y}$ is a fixed point if its components satisfy the equations

$$1 = \lambda_\nu p_\nu(y_\nu) \tag{9.9.19}$$

for $\nu = 1, 2, \ldots, k$.

In particular, if the survival probabilities has the Weibull form

$$p_\nu\left(y_\nu\right) = \exp\left(-\left(\beta_\nu y_\nu\right)^{\alpha_\nu}\right) \tag{9.9.20}$$

for every ν, then the components of a positive fixed point of equation (9.9.17) have the form

$$y_\nu = \frac{\left(\ln \lambda_\nu\right)^{\frac{1}{\alpha_\nu}}}{\beta_\nu}, \tag{9.9.21}$$

for every type $\nu = 1, 2, \ldots, k$, see section 9.5 for more details. Note $y_\nu > 0$ if, and only if, $\lambda_\nu > 1$. For this simple case, the results presented in section 9.5 may be used to determine parameter values such that the vector $\boldsymbol{y}$ with components in (9.9.22) is attracting. Although the case just considered is a very special case, it will, nevertheless, be useful in determining preliminary values of parameters such that a fixed point $\boldsymbol{y}$ is attracting when there is mutation among types and the functions $\gamma_\nu\left(\boldsymbol{y}\right)$ have the form $\gamma_\nu\left(\boldsymbol{y}\right) = y_1 + y_2 + \cdots + y_k$ for all types ν. In a subsequent chapter, the problem of finding parameter values such a fixed point is attracting will be studied in greater depth. Just as with the single type case studied in sections 9.4 and 9.5, the extent to which a solution of the embedded deterministic model is a good measure of central tendency for the sample functions of the stochastic process may be determined experimentally in Monte Carlo simulation experiments.

9.10 Estimating the Distribution of the Most Recent Common Ancestor in Simulated Genealogies

In the foregoing sections of this chapter, procedures for finding the most recent common ancestor of two individuals selected at random in a simulated genealogy were described for cases of one or more types. In this section, the results of two computer simulation experiments will be reported in which a one type self-regulating branching process was used to simulate genealogies derived from a single initial individual. In both these experiments, a Weibull survival function was used in the formulation of the self-regulating one type branching process and a Poisson was chosen for the distribution of the number of offspring produced by each individual. In order to avoid the problem in simulating a genealogy when a line becomes extinct, the process was modified by a procedure described in section 9.3 which assured that every individual in every generation would contribute at least one offspring to the next generation. The exercise of implementing a self-regulating branching process when the survival probability in (9.4.12) is in force will be left to the reader.

The one type self-regulating branching process considered in this section depends on four parameters; namely, α and β for the Weibull survival function, λ, the expected number of offspring produced by each individual and y_0, the size of the initial population. In experiment 1, the parameters α and β for the Weibull survival function were assigned the values $\alpha = 2$ and $\beta = 10^{-4}$, and the parameter λ for the Poisson distribution was assigned the value $\lambda = 0.05$. Because each individual produced at least one offspring, the expected number of offspring produced by each individual was $\lambda_1 = 1 + \lambda = 1.05$. Finally, the initial size of the population was chosen as $y_0 = 1$. Starting from one initial individual, a genealogy was simulated for 500 generations and the individuals in this genealogy were labeled following the procedures described in section 9.2. Altogether, this genealogy contained 5,659,445 descendants of the initial individual, and, in generation 500, the population contained 21,444 individuals. To estimate the distribution of the number of generations back in time to find the most recent common ancestor of two random selected individuals in generation 500, the procedure for calculating realizations of the random variable B described in section 9.2 was followed. The time taken to complete this experiment on a desk top computer operating at 3 gigahertz was about 4 hours. The reason for this relatively long execution time was that after the size of the population in any generation had grown to 21,000 or more, it was necessary to simulate over 21,000 realizations of a Poisson random variable in subsequent generations, which was time consuming. Presented in Figure 9.10.1 is a bar graph of the estimated distribution back to find the most recent common ancestor based on the random sampling of 1,000 pairs of individuals in generation 500.

Altogether there are 49 bars in this figure with base numbers ranging from 330 to 481 generations back in time to the most recent common ancestor of two randomly selected individuals. Most of the bars in the left most bars in the figure represent one to ten pairs of individuals. For example, in the bar representing 330 generations back, there was only one pair of individuals and the bar for 414 generations back contains 5 individuals. The modal or most frequent bar was 481 generations back which represented 506 pairs of individuals out of a total of 1000 pairs. The next most frequent classes back in time were 457, 461 and 476 generations with 89, 100 and 80 pairs of individuals, respectively.

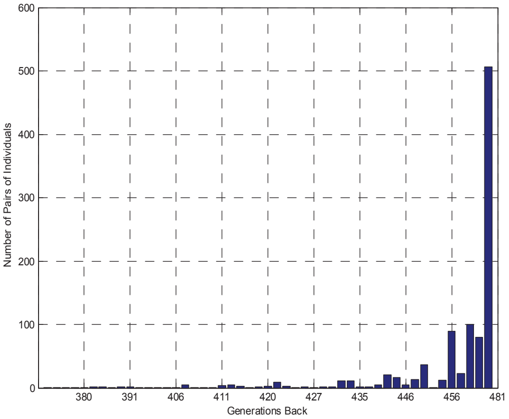

Figure 9.10.1 Estimated Distribution of Generations Back to the Most Recent Common Ancestor - Experiment 1.

The parameter assignments for experiment 2 differed from those in experiment 1 with respect to only two parameters. Thus, the parameter β in the Weibull survival function was assigned the value $\beta = 10^{-3}$ and the parameter λ was assigned the values $\lambda = 0.15$ so that the expected number of offspring contributed to the next generation by each individual was $\lambda_1 = 1 + \lambda = 1.15$. In this experiment, starting with one initial individual, a genealogy was computed for 1,000 generations. At generation 1000, there were 2,738 individuals in the population, but altogether, 2,339,836 descendants of the initial individual had been simulated during the 1000 generations under consideration. The time taken to complete this experiment was a little over one hour on a desk top computer running at 3 gigahertz. The key factor in this short execution time, when compared with experiment 1, was that the parameter β was assigned the value $\beta = 10^{-3}$ which lead to smaller population sizes and the need to compute fewer realization of a Poisson random variable once population size had grown to about 2,700 individuals. On the other hand, in experiment 1, where β had the value $\beta = 10^{-4}$, over 21,000 realizations of a Poisson random had to be computed in every generation once population size had reached about 21,000 individuals. Presented in Figure 9.10.2 is the estimated distribution of generation back in time to find the most recent common ancestor based on a sample size of 1,000 pairs of randomly selected of individuals.

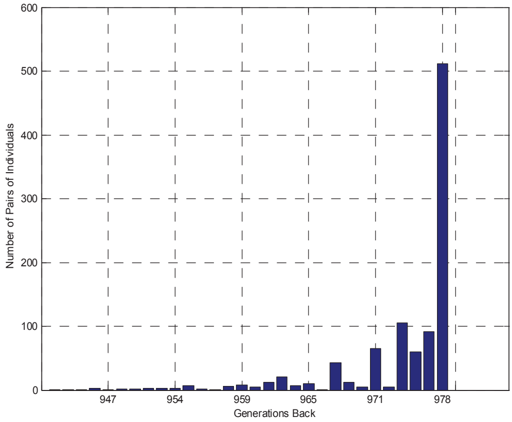

Figure 9.10.2 Estimated Distribution of Generations Back to the Most Recent Common Ancestor - Experiment 2.

The appearance of this figure is similar to Figure 9.10.1, but it has only 30 bars, ranging in base values form 942 to 981, rather than 49 bars in Figure 9.10.1. Like that in Figure 9.10.1, the distribution is skewed to the right, where the most frequent bar or classes occur. The most frequent bar or class with 512 pairs of individuals was 981 generations back to the most recent common ancestor. The next most frequent class was 978 generations back with 92 pairs of individuals. In other words, over 600 of the 1,000 pairs of individuals tested had as common ancestors individuals who were present in the population in generations 19 to 22. Even though the number of offspring produced by each individual in a simulated genealogy was not kept track of in the software, one suspects that the pairs of individuals in the most frequent classes were descendants of individuals in generation 19 to 22 who produced 2 or more offspring purely by chance. On the other hand, the bar for 942 generations back, had only one pair of individuals, who had a common ancestor in generation 58 of the simulated genealogy. Thus, it is tempting to conclude that such pairs of individuals were descendants of individuals who produced only one offspring.

With regard to the random number generators used in these experiments, it should be mentioned that the generator described in (5.7.11) was in force while simulating the genealogies, but in the experiments in which pairs of individuals were selected at random to find the most recent common ancestor, the generator in (5.7.4) was used, which is the default generator in the programming language *APL* 2000. The choice of the random number generator in (5.7.11) was a judicious choice for simulating genealogies, because large numbers of realizations of uniform random variables on the interval $(0, 1)$ were required in these simulations, which in turn required a generator with a long period. On the other hand, because procedures for drawing sample of $r < n$ objects from a set of n objects without replacement is preprogrammed in *APL*, using the generator in (5.7.4), the code designed to select 1,000 or more pairs of individuals at random could be written in a few lines, which greatly shortened the time required to write the software to execute the experiments reported in this section. Furthermore, because relatively few random numbers were required to select 1,000 pairs of individuals at random, the generator in (5.7.4) was judged to be adequate for this task. See chapter 5 for more details.

9.11 Connections Between the Embedded Deterministic Model and the Branching Process

In this section, the results of three computer experiments will be reported in which the trajectory computed by using the embedded deterministic model was compared with selected quantile and *MEAN* trajectories estimated from a sample Monte Carlo realizations of the self-regulating one type branching process described in previous sections. Throughout these experiments the Weibull survival function was used in the formulation of the self-regulating branching process with assigned parameter values $\alpha = 2$ and $\beta = 10^{-4}$. With this assignment of parameter values, the fixed point in (9.5.2) is attracting if, and only if, the parameter λ satisfies the condition $1 < \lambda < e = 2.718\,281\,828\,459\,05$, see (9.5.7). Thus, if λ is assigned the value $\lambda = 2.5$ then the fixed point, given by (9.5.2), is attracting, and, for this assignment of parameter values, the fixed point has the value

$$y_f = \frac{(\ln \lambda)^{\frac{1}{\alpha}}}{\beta} = 9572.3076. \tag{9.11.1}$$

Therefore, starting from any positive initial value, if the trajectory of the embedded deterministic model is computed recursively, using equation (9.4.7), then it would be expected to converge to this fixed point.

In experiment 1, the parameter λ was assigned the values $\lambda = 2.5$, the size of the initial population was chosen as $y_0 = 1$, the number of generations considered in the experiment was 1,000 and the quantile trajectories of the self-regulating branching process were estimated from a sample of 100

Monte Carlo realizations of the process. The trajectory based on the embedded deterministic model was also computed and graphed along with the Q_{25}, Q_{50} and Q_{75} quantile trajectories. The results of this experiment are graphically summarized in Figure 9.11.1 for 100 generations of a projection of 1000 generations.

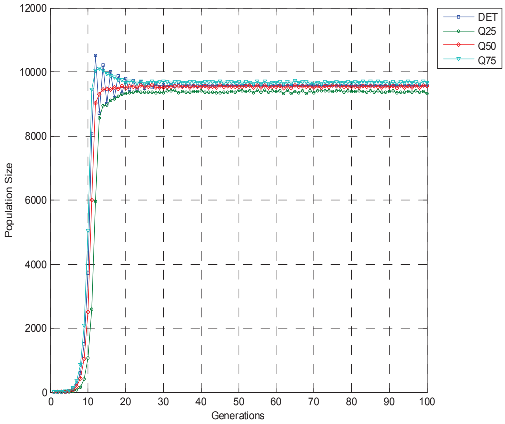

Figure 9.11.1 Quantile and Deterministic Trajectories When the Fixed Point is Attracting.

As can be seen from this figure, at about 10 generations into the projection, the trajectory of the embedded deterministic model was beginning to converge to the fixed point and the quantile trajectories of the process were also reaching values near the fixed point. If one is thinking about the evolution of mitochondrial DNA in a human population, 10 generations would correspond to about 150 years of evolution when the age of first child bearing for females is assumed to be about 15 years. The reason for choosing only 100 generations to display in a figure was that if 1,000 generations were represented in one graph, the transitory period of the first 10 generations of the projection was completely masked, resulting in an uninformative set of graphs. Among the 100 realizations of the process, there were a few cases in which the population became extinct, which is distinguishing property of the branching process. From the point of view of stochastic processes, the branching process under consideration is a discrete time Markov chain whose state space is the set of non-negative integers with 0 as the only absorbing state. Consequently, the clustering of the quantile trajectories process around the trajectory of the embedded deterministic model from generation 30 onwards may be interpreted as sampling from the quasi-stationary distribution of the branching process, given that extinction has not occurred. When a fixed point of the embedded deterministic model is attracting, the quasi-stationary distribution behavior observed in Figure 9.11.1 is quite typical of that which would have been observed if other parameters assignments had been used in the experiment.

In experiment 2, the parameter λ was assigned the value $\lambda = 3$ and the other input parameters had the same values as in experiment 1. With this assigned value for the parameter λ, the fixed point in (9.11.1) is no longer attracting so that it would be expected that the embedded deterministic model would behave differently from that in experiment 1. For with this assignment of parameters values, the embedded deterministic model converged to a stable two cycle trajectory in which population size fluctuated systematically among the two values of about, 7,411 and 12,857. Displayed in Figure 9.11.2, is the trajectory of the embedded deterministic model for the parameter assignments in experiment 2. As can be seen from this figure, after about 10 generations, the trajectory of the deterministic model converged to a stable two cycle trajectory in which population size fluctuated between two values ranging from less that 8,000 to over 12,000. Presented in Figure 9.11.3 are graphs of the $MEAN$ and $MEDIAN$, Q_{50}, trajectories based on a sample of 100 Monte Carlo realizations of the branching process. Interestingly, as can be seen from this figure, the Q_{50} trajectory behaves in a manner very similar to that of the deterministic trajectory in Figure 9.11.2 after about 20 generations. But, as expected, the fluctuation of the $MEAN$ trajectory was much less than that of the $MEDIAN$. Among the several cases investigated, but are not reported here, this case was selected for inclusion, because the $MEDIAN$ trajectory of the process behaved in a way that was very similar to that of the embedded deterministic model.

In experiment 3, the parameter λ was assigned the value $\lambda = 5$ so that, on average, every individual in any generation contributed 5 offspring to the next generation. Just as in experiments 1 and 2, all other parameters of the formulation under consideration had the same values as in experiment 1. In this experiment, however, the trajectory of the embedded deterministic model did not converge to a single value or a stable cycle of some order, but fluctuated wildly among population sizes, ranging from less than 5,000 to over 20,000. Displayed is Figure 9.11.4 is a statistical summarization of a sample of 100 Monte Carlo realizations of the branching process along with the irregular trajectory of the embedded deterministic model. The most

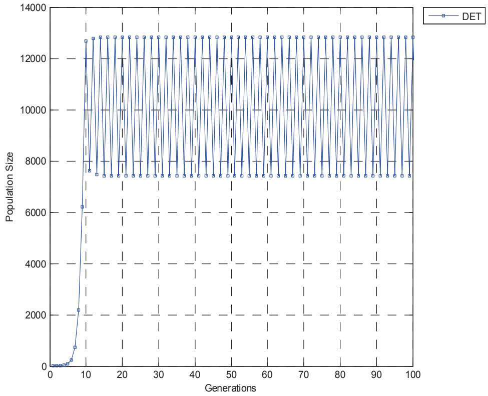

Figure 9.11.2 Convergence of Deterministic Model to a Stable Two Cycle Trajectory.

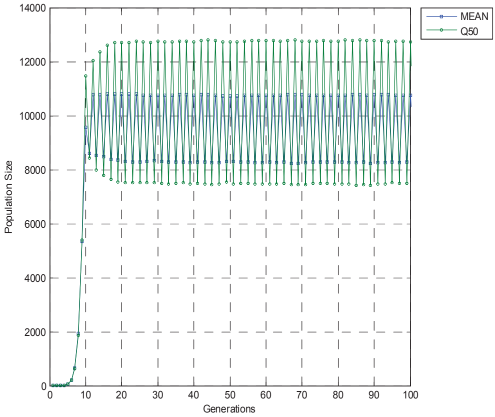

Figure 9.11.3 MEAN and *MEDIAN* Trajectories for Stable Two Cycle Model.

striking property of this figure is that the Q_{25}, Q_{50} and Q_{75} trajectories fluctuate much less from generation to generation than those of the deterministic model after 10 or more generations into the sample of projections. As can be seen from the Q_{50} trajectory, population size in 50 percent of the realizations of the branching process were below about 7,500 individuals. Similarly, from an inspection of the Q_{75} trajectory, it can be seen that 25 percent of the realizations of the process are near or above a population size of 20,000 individuals. The relatively smooth quantile trajectories of the process may be interpreted as those of a stochastic process that is in a quasi-stationary statistical equilibrium, given that extinction of the process has not occurred. Unlike the projection based on the embedded deterministic model, in which the population never becomes extinct, in the process extinction may occur with positive probability. Actually, in a sample of 100 realizations of the process for 1,000 generations, only in one realization did the population become extinct, which is consistent with empirical observations made by biologist that among all species that have existed on the earth during its evolution, most have become extinct at some time in the past so that in evolutionary time extinction is a common event.

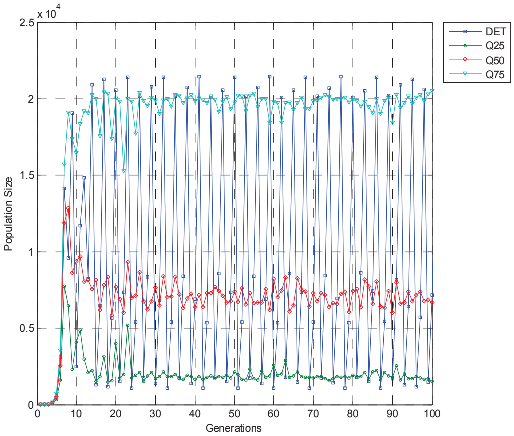

Figure 9.11.4 Selected Quantiles of the Process with a Chaotic Trajectory of the Deterministic Model.

The assigned parameters values for experiment 3 gave rise to a model for the evolution of a population that is characterized by large fluctuations in population size from generation to generation. Examples of such populations may be found among insect species in which fluctuations in population size among generations may vary by orders of magnitude. The self-regulating branching process under consideration is useful in attempts to model such evolutionary process, because the magnitude of such fluctuations is determined by the small values of the parameter β. For example, in an experiment not reported here, the parameter β was assigned the value

$\beta = 10^{-7}$, which lead to population sizes varying among thousands to millions of individuals among generations.

It should be mentioned that the choices of values for the parameter λ used in the experiments reported in this section were not an accident but were due to some prior knowledge of a field of mathematics called chaos theory. For if one were to plot the formula

$$f(y) = \lambda y \exp\left(-\left(\beta y\right)^{\alpha}\right) \tag{9.11.2}$$

for the embedded deterministic model displayed in (9.4.7) as a function of population size $y \geq 0$ for fixed parameter values $\alpha = 2$ and $\beta = 10^{-4}$ and any fixed value of the parameter λ, then starting form $y = 0$ the graph would rise from 0 to a maximum and then decline to 0 as $y \uparrow \infty$. As is well known from chaos theory, such a graph is the signature of a model that will lead to chaos when the function in (9.11.2) is iterated starting from an initial point y_0. That is, given y_0, y_1 is determined by $y_1 = f(y_0)$ and so on until all desired values in a sequence $y_0, y_1, y_2, \ldots$ are computed recursively.

Furthermore, if one were to calculate the so called orbit diagram, it would reveal how for fixed values of the parameters α and β, one would proceed from points of attraction to a fixed point, to stable two cycles, to stable three cycles and so on to what is called chaos as the parameter λ moves from values slightly greater than one to larger values. If a reader is interested, it is suggested that the key phrase "orbit diagram" be typed into a search engine for the world wide web where this mathematical phenomenon will be discussed. One may also consult books on chaos. It was prior knowledge of this type that lead to choosing the three values λ used in the experiments reported above. It should also be mentioned that the function in (9.11.2) was derived from a form of plausible of reasoning that is widely used in formulations of stochastic models and not from a desire to deal with a model that may lead to chaos. Nevertheless, in subsequent chapters, whenever a deterministic model embedded in a stochastic process arises, the behavior of the sample functions of the process and the trajectory of the deterministic model will be compared whenever it is stable or chaotic.

9.12 Procedures for Simulating Realizations of a Poisson Random Variable

In this and previous chapters, realizations of Poisson random variables were computed in a number of Monte Carlo simulation experiments, but no details were given regarding the procedures used to compute these realizations. In this section, several procedures for computing realizations of a Poisson random variable will be outlined. One procedure, which works well for small λ, makes use of a connection between a random variable X with an exponential distribution with scale parameter $\lambda > 0$ and a Poisson random variable Y with expectation λ. Let $X_1, X_2, \ldots$ be a sequence of independent exponential random variables with scale parameter λ. Then, for every integer $y \geq 1$, it follows that

$$P\left[Y = y\right] = P\left[X_1 + X_2 + \cdots + X_y \leq 1 < X_1 + \cdots + X_{y+1}\right]. \quad (9.12.1)$$

Algorithm Q on page 223 of Kennedy and Gentle (1980), which is an implementation of the ideas in (9.12.1), was implemented and used to compute realizations of Poisson random variables when λ was small from realizations of uniform random variable on the interval $[0, 1) = [x \mid 0 \leq x < 1]$. Because λ was small in all experiments in which genealogies were simulated, this algorithm worked well in the sense that genealogies consisting of 500 to 1,000 generations could be simulated in acceptable periods of time on desk top computers.

In those cases in which a Poisson approximation to the binomial was used, however, the parameter λ may be too large to use the above algorithm efficiently. For those cases a more general algorithm was used to simulate realizations of a Poisson random variable with a parameter λ. Even though the support of a random variable with a Poisson distribution is the infinite set of non-negative integers, $(0, 1, 2, \ldots)$, its essential support consists of some finite set of integers $(0, 1, 2, \ldots, m)$, where m is an integer such that relation

$$\sum_{y=0}^{m} \exp\left(-\lambda\right) \frac{\lambda^y}{y!} \simeq 1 \quad (9.12.2)$$

holds approximately on the computer an investigator may be using. Given this value of m, a good finite approximation to the Poisson distribution would be a density of finite support defined by the equation

$$f\left(y\right) = \frac{1}{\sum_{y=0}^{m} \exp\left(-\lambda\right) \frac{\lambda^y}{y!}} \exp\left(-\lambda\right) \frac{\lambda^y}{y!} \quad (9.12.3)$$

for $y = 0, 1, 2, \ldots, m$. The distribution function corresponding to this density is

$$P\left[Y \leq y\right] = F\left(y\right) = \sum_{\nu=0}^{y} f\left(\nu\right), \quad (9.12.4)$$

and is defined for $y = 0, 1, 2, \ldots, m$.

To simulate a realization of a random variable Y with this distribution function, let u denote a realization of a uniform random variable U on the interval $[0, 1)$. Then, a realization of Y may be computed by finding the largest integer y in the set $[\nu \mid \nu = 0, 1, 2, \ldots, m]$ such that $F(y) \leq u$. Finding the number y is relatively easy for those programming languages that have the property of generating Boolean indicators consisting of ones and zeroes, depending on whether some logical relationship concerning numbers is true or false. For example, let

$$\boldsymbol{F} = (F(0), F(1), \ldots, F(m)) \tag{9.12.5}$$

denote an array of non-decreasing numbers computed from the distribution function in (9.12.4). Then, just as in section 9.2, the logical relationship

$$\boldsymbol{F} \leq u \tag{9.12.6}$$

would generate an array of Boolean indicators $(\xi_0, \xi_1, \ldots, \xi_m)$ such that $\xi_\nu = 1$ if $F(\nu) \leq u$ and $\xi_\nu = 0$ otherwise for $\nu = 0, 1, 2, \ldots, m$. Now consider the sum

$$\sum_{\nu=0}^{m} \xi_\nu. \tag{9.12.7}$$

The range of this sum is the set of integers $(1, 2, \ldots, m+1)$. For if all the Boolean indicators are one, then this sum has the value $m + 1$. Also note that the Boolean indicator for the occurrence of the event $F(0) \leq u$ is one and not zero even if 0 is the largest integer such that this inequality is true. Therefore, to find the greatest integer y is the set $(0, 1, \ldots . m)$ such that $F(y) \leq u$, one must be subtracted from the sum in (9.12.7). Thus, the integer y is determined by the expression

$$y = \left(\sum_{\nu=0}^{m} \xi_\nu \right) - 1. \tag{9.12.8}$$

Briefly, whenever λ had value such that the use of algorithm Q would not be efficient, the ideas just outlined were used to write computer code to simulate realizations of a Poisson random variable with some expectation λ. The precise form of the code designed to implement these ideas would depend on the properties of programming language used to write the code. The performance of the code was always checked in preliminary computer experiments in order to assess whether realized values of Poisson random variables followed well known laws of probability. For example, let $y_1, y_2, \ldots, y_n$ be a large sample of realizations of a Poisson random variable with expectation λ. Then, by the law of large numbers, the sample $MEAN$ $\overline{y} = (y_1 + y_2 + \cdots + y_n)/n$ should be close to λ. In many experiments, the $MEAN$ was indeed close to λ for sample sizes of 10,000 or more in every implementation of the ideas outlined in this section to simulate realizations of Poisson random variables in the computer experiments reported in this and other chapters. Other tests were also sometimes performed. Among these tests was that estimating the density function of the random variable under consideration from simulated data and comparing it to that used to simulate the data. In all such cases considered, the software also passed tests of this type.

As was shown in section 9.3, when simulating realizations of a branching process, one encounters random sums of the form

$$Y_n = \sum_{\nu=1}^{Y_{n-1}} X_\nu, \tag{9.12.9}$$

where Y_n is the size of the population in generation $n \geq 1$ and, given $Y_{n-1} = y_{n-1}$, the random variables $X_1, \ldots, X_{y_{n-1}}$ are conditionally independent Poisson random variables with a common expectation λ. Whenever it is desired to simulate genealogies, a realization of each random variable in the sum in (9.12.9) must be computed, but, when attention is focused only on total population size, computing only realizations of the sum in (9.12.9) will suffice. As it turns out, when $Y_{n-1} = y_{n-1}$ is large, then computing a realization of the sum Y_n can be reduced to calling only one realization of a standard normal random by an application of the central limit theorem.

To verify this statement, observe that by the central limit theorem, given that $Y_{n-1} = y_{n-1}$ is large, the random variable

$$\frac{Y_n - y_{n-1}\lambda}{\sqrt{y_{n-1}\lambda}} \simeq Z \tag{9.12.10}$$

is approximately normally distributed in the sense that the random variable Z is normal with expectation 0 and variance 1. Therefore,

$$Y_n \simeq y_{n-1}\lambda + \sqrt{y_{n-1}\lambda} \times Z. \tag{9.12.11}$$

Consequently, to simulate a realization of Y_n using this formula, it suffices to call for one realization of a standard normal random variable Z and adjust the result so that Y_n is a positive integer.

In experiments 1 and 2 reported in section 9.11, the Q algorithm was used whenever $y_{n-1} < 2,000$ and when $y_{n-1} \geq 2,000$ the normal approximation in (9.12.11) as used. But in experiment 5 reported in section 9.11

where the parameter λ had the value $\lambda = 5$, the Q algorithm was used only when $y_{n-1} < 1,000$ and the normal approximation was used when $y_{n-1} \geq 1,000$. Using the normal approximation in the experiments reported in section 9.11, greatly reduced the among of computer time needed to simulate 1,000 generations of self-regulating branching process.

The Q algorithm was designed specifically to simulate realizations of a Poisson random variable when the expectation parameter λ is small. The ideas underlying the procedures used to simulate realizations of a Poisson random variable when λ is large so that the Q algorithm cannot be used efficiently are, in fact, quite general and may be used as a basis for writing computer code when it is desired to simulate realizations of distributions of random variables other than those of the Poisson family. For example, any distribution an investigator wishes to entertain with the set of integers $(0, 1, 2, \ldots, m)$ as its support may be used in place of the Poisson type density in (9.12.3). Furthermore, if λ and σ^2 are the expectation and variance of this distribution, then the central limit theorem may be applied to implement an approximation similar to that in (9.12.11) for simulating realizations of sums of random variables.

Bibliography

[1] Asmussen, S. and Hering, H. (1983) **Branching Processes**. Birkhauser, Boston, Basel, and Stuttgart.

[2] Athreya, K. B. and Ney, P. (1972) **Branching Processes**. Springer-Verlag, Berlin, Heidelberg, New York.

[3] Durrett, R. (2008) **Probability Models for DNA Sequence Evolution, Second Edition.** Springer Science.

[4] Gulick, D. (1992) **Encounters With Chaos**. McGraw-Hill, New York, London.

[5] Haccou, P., Jagers, P. and Vatutin, V. A. (2007) **Branching Processes - Variation, Growth and Extinction of Populations**. Cambridge University Press, The Edinburgh Building, Cambridge, CB2 8Ru, UK.

[6] Harris, T. E. (1963) **The Theory of Branching Processes**. Springer-Verlag, Berlin, Heidelberg, New York.

[7] Jagers, P. (1975) **Branching Processes with Biological Applications.** John Wiley and Sons, London, New York, Sydney, Toronto.

[8] Jagers, P. and Sagitov, S. (2004) Convergence to the Coalescence in Populations of Substantially Varying Size. Jour. Appl. Prob. **41**:368–378.

[9] Kemeny, J. G., Snell, J. L. and Knapp, A. W. (1966) **Denumerable Markov Chains**. D. Van Nostrand Inc., Princeton, Toronto, New York, London.

[10] Kennedy, W. J. and Gentle, J. E. (1980) **Statistical Computing**. Marcel Dekker, Inc. New York.

[11] Kingman, J. F. C. (1982a) On the Genealogy of Large Populations. In **Essays in Statistical Science**. Journal of Applied Probability, Special Volume 19A.

[12] Kingman, J. F. C. (1982b) The Coalescent. **Stochastic Processes and Their Applications**, **13**:235–248

[13] Mode, C. J. (1971) **Multitype Branching Processes - Theory and Applications**. American Elsevier, New York, Toronto.

[14] Nordborg, M. (2001) Coalescent Theory in **Handbook of Statistical Genetics,** pp. 723–765, Wiley, New York.

[15] Pollak, E. (2007) Coalescent Theory for a Completely Random Mating Population. Math. Biosci. 205:315–324.

Chapter 10

Emergence, Survival and Extinction of Mutant Types in Populations of Self Replicating Individuals Evolving From Small Founder Populations

10.1 Introduction

There is an extensive literature on conceptual issues underlying evolutionary biology. An example of a recent book dealing with this topic is that edited by Sober (2006), which contains a collection of essays by many authors on such topics as fitness, units of selection as well as many other topics that have arisen in the ongoing process of developing the conceptual foundations of Darwinian evolution. In this connection, there is also the recent book by Okasha (2006), in which ideas concerning evolution and levels of selection are examined critically. Another book in a similar critical vein is that of Pigliucci and Kaplan (2006) in which the conceptual foundations of evolutionary biology are examined. It is beyond the scope of this chapter to review the concepts treated in these books, and, in what follows, attention will be confined to studying the emergence, survival and extinction of mutations within the conceptual framework of the class of self regulating mulitype branching processes introduced in chapter 9.

When considering the evolution of populations evolving according to the concepts of Darwinian evolution, it is necessary to entertain models with at least two types of individuals, for if a population is composed of individuals of only one type, then it would not contain the necessary variability upon which natural or artificial selection acts to drive the evolution of the population. Similarly, Darwinian evolution can not act on a population if there is no process or mechanism, such as mutation, that accounts for the emergence of new variability in a population. The processes underlying mutations, which give rise to variation in a population, are varied and complex as discussed briefly in chapter 8. Therefore, for the sake of simplicity, the model of mutation introduced in chapter 9 in connection with the formulation of multitype branching processes will again be utilized in this chapter. As alluded to above, selection is also among the driving forces of evolution, which may in turn consist of many components. In this chapter, however, only two components of natural selection will be considered, which can be formulated explicitly within a framework of self regulating multitype branching processes. These two components will be designated as reproductive capacity of individuals and their ability to compete with other individuals in a population.

The idea of reproductive capacity of an individual of a given type may be quantified in terms of the λ-parameters, representing the expected number of offspring produced by each type of individual in any generation. Specifically, for the sake of simplicity, only three types of individuals will be considered throughout this chapter so that in any computer experiment a vector with three parameters $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)$ must be assigned positive numerical values. It is of interest to note that this assignment of parameter values is consistent with the definition of fitness proposed in an essay by Mills and Beatty in the book Sober (2006) cited above. The ability of individuals to compete with others in a population will be characterized in terms of the parameters in the Weibull survival function introduced in chapter 9 in connection with the formulation of self regulating multitype branching processes. For fixed values of the parameters $\alpha_\nu > 0$ for $\nu = 1, 2, 3$, the ability of types of individuals to compete with others will be quantified by assigning positive values to the components in the vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$. According to this formulation, the smaller the value of a β-parameter for a given type, the greater is the ability of an individual of this type to compete with others in a population. For example, if these parameters were assinged the values $\beta_1 = 10^{-5}, \beta_2 = 10^{-6}$ and $\beta_3 = 10^{-7}$, then individuals of type 2 are more competitive than those of type 1 and individuals of type 3 are more competitive than either those of types 1 or 2.

Like all proposed models for the evolution of populations, application of the self regulating branching processes under consideration will be most readily applicable to the following types of populations. Among these types are populations consisting of macro organic molecules with a capacity to mutate, as described in a series of interesting lectures on the origins of life by Hazen (2005), and with capabilities for self replication. Such macro molecules include DNA and RNA. Other types of populations may include haploid types of inheritance characterized by the passage of mitochondrial DNA through mother's lines of descent and that on a segment of the Y

chromosome in males, passed through males lines of descent, that does not undergo genetic recombination during meiosis. When applying the theory with respect to haploid inheritance, one should be mindful of the possibilities of interactions of genes in these haploid segments of DNA with other genes in the nuclear and mitochondrial genomes.

A third type of population to which the theory may be applicable are those consisting of monoecious individuals that are such that each individual has both female and male reproductive organs and is, therefore, capable of producing offspring that arise through the union of female and male gametes of the same individual. Thus, reproduction in such populations does not involve the interactions of female and male individuals. Examples of populations of this type are those of common bread wheat, which reproduces by a process called selfing. However, even though the interaction of two sexes is not involved in the evolution of such populations, individual plants compete for nutrients in the soil and light from the sun, and, thus competition is one of the driving forces of evolution in such populations.

It should be mentioned that in order to study the evolution of populations of monoecious diploid individuals, it would be necessary to reformulate the mutation matrix $\mathfrak{M}$ of the form displayed in (9.8.19). For example, suppose one locus with two alleles, A and a is under consideration. Then, the three genotypes, AA, Aa and aa would be the three types under consideration in a multitype branching process. If it is also assumed that both alleles A and a may mutate to the other, then one could introduce a mutation matrix

$$\mathbf{\Delta} = \begin{pmatrix} \delta_{11} & \delta_{12} \\ \delta_{21} & \delta_{22} \end{pmatrix}, \tag{10.1.1}$$

where every element δ_{ij} in the matrix satisfies the condition $\delta_{ij} \geq 0$ and each row sums to 1. By way of illustration, δ_{12} is the probability per generation that allele A mutates to a, and δ_{11} is the probability that no mutation of A occurs.

Let the three genotypes under consideration, AA, Aa and aa, be indexed by the numbers 1,2,3, and let π_{ij} denote the probability that a genotype of type i produces an offspring of genotype j. Then, the next step in the formulation of the model would be that of deriving formulas for the elements of the 3×3 matrix

$$\mathbf{\Pi} = (\pi_{ij})\,. \tag{10.1.2}$$

Briefly, the rationale used to derive formulas for the elements of this matrix would be similar to that used in the deriving the elements in table 5.5.3 for the model of inherited autism developed in chapter 5, but the derivation of these formulas will be left as an exercise for the reader. Given the matrix $\mathbf{\Pi}$, the evolution of the population could be expressed in terms of a self regulating multitype branching process similar to that discussed in chapter 9 by letting the matrix $\mathbf{\Pi}$ play the same role as the matrix $\mathfrak{M}$ in the formulation of the multitype branching process described in chapter 9. Although the model just discussed is of considerable interest, it was not implemented or used in the computer experiments reported in the sections that follow.

In a word, the computer experiments presented in this chapter are extensions of the discussion on the survival of mutant genes described by Fisher (1958) in his well known book on the genetical theory of natural selection. Unlike the simple model utilized by Fisher however, which is now known as a one type Galton-Watson process with a Poisson offspring distribution, it is currently feasible to use modern technology consisting of fast desk top computers and user friendly software to conduct computer simulation experiments based on branching processes that were inconceivable during the life span of Fisher, one of the founders of the subject of mathematical genetics. A theme common to all experiments presented in the sections that follow is that a population evolves from an initial founder population consisting of a small number of individuals. For those readers who are interested in further applications of branching processes in biology and the mathematics underlying this class of stochastic processes, it is recommended that the recent book by Haccou, Jagers and Vatutin (2007) be consulted.

As will be observed in the computer experiments reported in the sections that follow, there is an apparent tendency for multitype self regulating branching processes to converge, within hundreds or thousands of generations, to what appears to be a stationary distribution. To account for this phenomenon, it is appropriate to discuss briefly the mathematical foundations underlying the class of stochastic processes studied in this chapter, which provide a basis for understanding this phenomenon. With the exception of the process which by assumption evolves in a random environment, all processes considered in this chapter are Markov chains whose state space is the set $\mathfrak{S}_k$ of all k-dimensional vectors of nonnegative integers of the form $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)$ for the case $k = 3$. Moreover, for all processes in this class, the state $\mathbf{0} = (0, 0, \ldots, 0)$ is an absorbing state and the set of states $\mathfrak{T} = \mathfrak{S}_k - (\mathbf{0})$ is the set of transient states for this class of processes. Although a formal proof that a quasi-stationary distribution with the set $\mathfrak{T}$ as its domain exists has not been given, the empirical evidence suggests that such a distribution does indeed exist for the class of Markov chains under consideration. The set $\mathfrak{S}_k$ is countably infinite and so is the set $\mathfrak{T}$ of transient states. However, if some finite subset of the set $\mathfrak{S}_k$ were chosen as the state space of the process, then the sort of argument outlined in chapter 4, where a formula for the quasi-stationary distribution was derived for absorbing Markov chains with a finite state space, could be invoked to prove that a quasi-stationary distribution did exist for this class of processes with a finite state space, see section 4.6 for details.

10.2 Experiments with the Evolution of Small Founder Populations with Mutation but no Selection

In this section, the evolution of a population with three types of individuals will be studied in computer simulation experiments when the size of the initial founder population was only 20 individuals of type 1. The objectives of these experiments were to investigate the survival of the population, and, if it survives, its evolution into a polymorphic population with three types of individuals. In experiment 10.2.1, the number of Monte Carlo replications chosen was 100, the simulation was carried out for 6,000 generations, and the probabilities of mutation among the types were assigned the constant $\mu_{ij} = 10^{-6}$ for $i \neq j$. To quantify the idea that there was no selection, the λ-parameters were assigned the common value $\lambda_i = 1.50$ for $i = 1, 2, 3$. Finally, the α and β parameters for the Weibull survival function were assigned the values $\alpha_i = 2$ and $\beta_i = 10^{-8}$ for $i = 1, 2, 3$. With these chosen values of the λ-parameters, the expected number of offspring produced by each individual, it was anticipated that the population would grow quite rapidly to a size such that mutant types would appear in the population with high probability. The symbolic form of the Weibull survival function used in this section for individuals of type ν, as well as in all sections of this chapter in generation n, was

$$p_\nu \left(\boldsymbol{y} \left(n \right) \right) = \exp \left(- \left(\beta_\nu y \left(n \right) \right)^{\alpha_\nu} \right), \tag{10.2.1}$$

where $\nu = 1, 2, 3$, $\boldsymbol{y}(n) = (y_1(n), y_2(n), y_3(n))$ is a vector denoting the number of individuals of each type present in the population in generation n and $y(n) = y_1(n) + y_2(n) + y_3(n)$ is total population size. Section 9.9 may be consulted for further details.

Presented in table 10.2.1 is a snapshot of the evolution of the population implied by the deterministic model embedded in the stochastic process at selected generations.

Table 10.2.1 Values of the Embedded Deterministic Trajectories for Experiment 10.2.1 at Selected Generations

Generation	Type 1	Type 2	Type 3
10	$1,153.278$	0.00003	0.00003
20	$66,502.45$	1.330064	1.330064
1000	$63,548,981.00$	$63,580.82$	63.580.82
3000	$63,295,799.00$	$190,171.70$	$190,171.70$
6000	$62,911,863.00$	$378,639.40$	$378,639.40$

From this table it can be seen that by generation 10, the number of type 1 individuals in the population had increased to an estimated 1,153 individuals, but the mutant types, 2 and 3, were very rare. By generation 1,000, however, the population consisted of over $63,000,000$ individuals of type 1 and about $63,000$ individuals each of the mutant types 2 and 3.

By generations 3,000 and 6,000, the number of individuals of type 1 in the population was still about $63,000,000$ individuals, but the numbers of individuals of mutant types 2 and 3 in these generations were about 190,000 and over 378,000, respectively.

These results show the even after 6,000 generations of evolution, the trajectories of the three types of individuals had not converged to constants and that the effects of the small founder population of only 20 individuals of type 1 persisted in that type 1 was still the predominant type of individual in the population. Before presenting the corresponding experiments with the three type stochastic model, it is appropriate to mention why the numbers of types 2 and 3 in table 10.2.1 are the same.

As it turned out, under the parameter assignments used in experiment 10.2.1 to simulate neutral evolution with mutation but no selection, the 3×3 matrix $\mathfrak{M}$ of mutation probabilities is symmetric and has the property that each row and column sums is 1. Such matrices are also sometimes called doubly stochastic and have the property that their stationary distribution is uniform. Given this result and the assumption that there is no selection, *i.e.*, all the λ-parameters were assumed to be equal, it can be shown that the numbers of types 2 and 3 in a projection determined by the deterministic model will always be equal, even though their numbers may increase from generation to generation. As will be seen in subsequent sections of this chapter, this property of equality will no longer hold when selection is introduced into assignments for parameters values.

Table 10.2.2 contains the Q50 trajectories as estimated from the sample 100 Monte Carlo realizations of the stochastic process with three types.

Table 10.2.2 Values of the Q50's Trajectories of the Stochastic Process for Experiment 10.2.1 at Selected Generations

Generation	**Type 1**	**Type 2**	**Type 3**
10	$1,094$	0	0
20	$62,807$	23	22
$1,000$	$63,434,625$	$103,564$	$90,589$
$2,000$	$63,171,239$	$233,614$	$230,339$
$3,000$	$62,784,385$	$433,027$	$431,122$

As can be seen from this table, by generation 10, the median number of individuals of type 1 in the population had risen to 1,094, but the Q50 trajectories for types 2 and 3 were zero, indicating that no mutations had appeared in the population by generation 10. By generation 20, however, mutant types 2 and 3 were present in the population in small numbers even though the median number of individuals of type 1 had risen to over 62,000. At generation 1,000, the Q50 trajectory for individuals of type 1 had risen to over $63,000,000$, which is a value similar to that produced by the embedded deterministic model. On the other hand, the Q50 trajectories for the mutant type 2 and 3 had risen to over 103,000 and 90,000, respectively, by generation 1,000. The Q50 trajectories for type 1 at generations 3,000 and 6,000 were still at about $63,000,000$, but the Q50 trajectories of types 2 and 3 continued to increase and over 230,000 and 400,000, respectively, in these generations.

By comparing tables 10.2.1 and 10.2.2, it can be seen that the embedded deterministic model consistently under estimated the number of mutant types in the population for the generations considered in these tables. Like that for the deterministic model, type 1, with an initial number of only 20 individuals, was the predominate type in the population according to the stochastic projection after 6,000 generations of evolution. Interestingly, experiment 10.2.1 demonstrates that the effect of a small number of initial individuals of one type can persists in the populations for a large number of generation when there is mutation but no selection. Unfortunately, for this experiment graphical presentation of the evolution of the population according to the deterministic and stochastic models were not informative, due to the rather rapid rate at which population increased.

In experiment 10.2.2, all the parameter values assigned in experiment 10.2.2 were still in force except that the λ-parameters were assigned the smaller values $\lambda_i = 1.01$ for $i = 1, 2, 3$, indicating that no type had a selective advantage over the other. Given these assignments of the λ-parameters, it was expected that the population would grow more slowly than in experiment 10.2,1 so that it would take more time before the population would reach a size sufficient to make the occurrence of mutations in the population more probable. Presented in table 10.2.3 are the trajectories of the embedded deterministic model at selected generations.

Table 10.2.3 Values of the Embedded Deterministic Trajectories for Experiment 10.2.2 at Selected Generations

Generation	Type 1	Type 2	Type 3
10	21.022	0.00021	0.00021
20	22.097	0.000442	0.000442
1,000	2,925.658	2.927	2.927
3,000	6,997,291.00	20,963.21	20,963.21
6,000	6,978,264.00	41,994.50	41,994.50

From this table it can be seen, as expected, that the growth of the population was much slower than for that in experiment 10.2.1, see table 10.2.1. For example, in generations 3,000 and 6,000 the estimated number of individuals of type 1 was only over 6,900,000 and the number of mutant individuals of types 2 and 2 were about 20,000 and 40,000, respectively, in these generations. As will be shown by looking at the statistical summaries of the Monte Carlo simulation experiment however, the embedded deterministic model for this experiment misrepresented what actually happened during the evolution of the stochastic process. Presented in table 10.2.4 are the *MAX* trajectories of the stochastic process as estimated from 100 Monte Carlo realizations of the process.

Table 10.2.4 Values of the Max Trajectories of the Stochastic Process for Experiment 10.2.2 at Selected Generations

Generation	Type 1	Type 2	Type 3
10	63	0	0
20	99	0	0
1,000	61,926	3,034	6,457
3,000	6,931,215	1,337,965	2,518,541
6,000	6,963,818	1,253,654	2,915,491

As indicated in the title of this table, only the values of *MAX* trajectories for the three types are presented at selected generations. The reason displaying only the *MAX* trajectories was that in the statistical summaries the $Q75$ trajectory was 0 at generation 6,000 and also in many preceding generations for each of the three types, indicating that for at least 75 percent of 100 simulated realizations of the process the population had become extinct within 6,000 or fewer generations. Consequently, the MAX trajectories represent those rather rare cases among the 100 realizations of the process in which the initial population of 20 individuals of type 1 grew to a population whose size was sufficient for mutant type to appear with high probability. Although the large amount of simulated data were not observed in detail due to difficulties in visually coping with large quantities of data, it is estimated that the MAX values in table 10.2.4 represent only 1 to 5 realizations of the process among the 100 that were simulated.

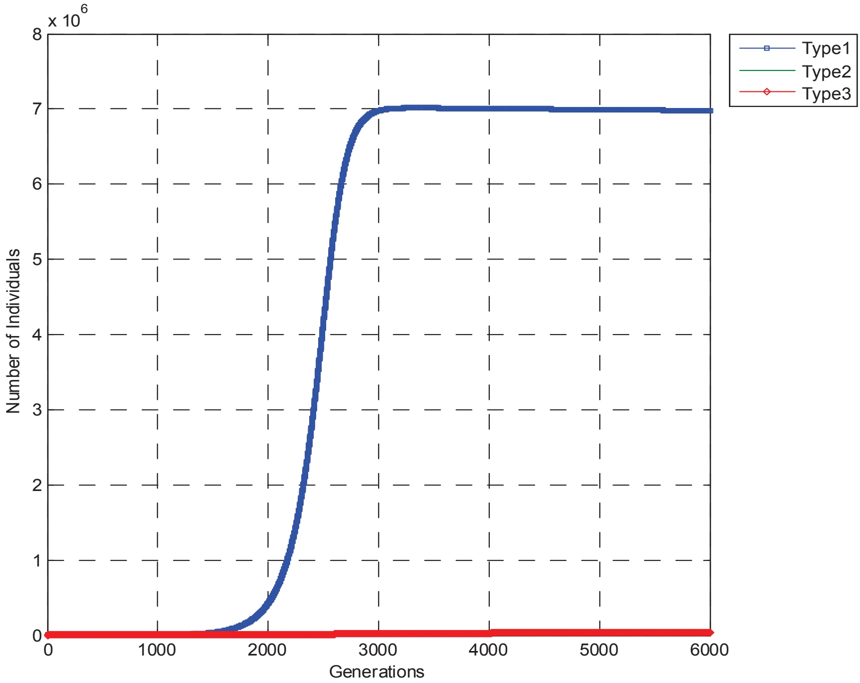

Figure 10.2.1 Trajectories of the Three Types for Experiment 10.2.2 as Computed From the Deterministic Model.

Even though tables 10.2.3 and 10.2.4 are useful in getting some ideas as the how the deterministic and stochastic process evolved, graphical representations of an experiment will, in general, provide a more informative overview of the results of simulation experiments. Presented in Figure 10.2.1 are the graphs of the trajectories of the three types as computed from the deterministic model embedded in the stochastic process. From this figure it can be seen that the number of individuals of type 1 in the population did not begin to rise significantly until about 2000 generations into the projection. Just as in experiment 10.2.1 however, type 1 individuals were the predominant type in the population, indicating that under slow population growth and neutral evolution with mutation the initial number of 20 individuals of type 1 had an effect that was still in force after 6,000 generations of evolution.

Figure 10.2.2 contains the MAX trajectories for the three type of individuals, which were computed from a sample of 100 realizations of the stochastic process in which it was estimated that only 1 to 5 realizations of the process did not become extinct.

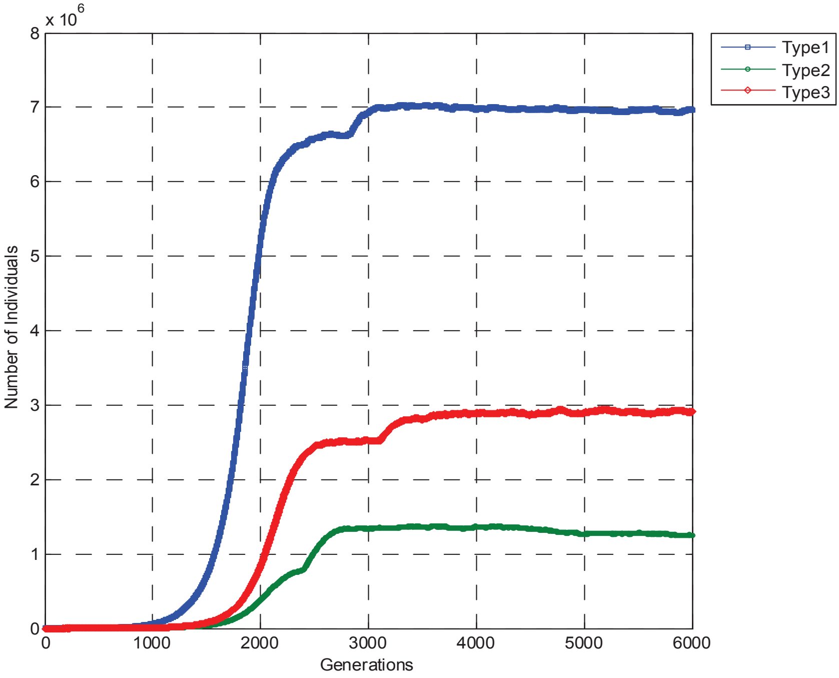

Figure 10.2.2 MAX Trajectories for the Three Types Based on Monte Carlo Simulated Data From Experiment 10.2.2.

From this figure it can be seen that, among those realizations, the MAX trajectories for each of the three types started to rise after about 1000 generations of evolution which were sooner than those for the deterministic projection. In this figure, it is interesting to note that, in relative terms, the MAX trajectories for types 2 and 3 are larger than those for the deterministic projection in figure 10.2.1. As neither types 2 or 3 had a selective advantage over the other, the observation that the MAX trajectory for type 3 exceeds that for type 2 throughout most of the projection is attributed to purely stochastic effects as observed in cursory inspections of the Monte Carlo simulated data in experiment 10.2.2.

10.3 Components of Selection – Reproductive and Competitive Advantages of Some Types

In this section, the embedded deterministic model and sample of Monte Carlo simulations of the three type stochastic process will be applied in an attempt to study some components of selection when there are mutations among the three types. Specifically, selection will be defined in terms of reproductive and competitive advantages of some types. In experiment 10.3.1, 100 Monte Carlo realizations of the stochastic process were computed for 6,000 generations, and, just as in the experiments discussed in section 10.2, the mutation probability among the three types were chosen as $\mu_{ij} = 10^{-6}$ if $i \neq j$.

To accommodate the idea that mutant types 2 and 3 had a reproductive advantage of over type 1, the parameter λ_1 was assigned the value $\lambda_1 = 1.005$ and the parameters for the two mutant types were assigned the value $\lambda_i = 1.01$ for $i = 2, 3$. Just as in the experiments of section 10.2, the α-parameters for the Weibull survival function were assigned the value $\alpha_i = 2$ for $i = 1, 2, 3$. To accommodate the idea that the three types differed in competitive abilities, the β-parameters were assigned the values $\beta_1 = 10^{-6}, \beta_2 = 10^{-7}, \beta_3 = 10^{-8}$, respectively, for types 1, 2 and 3. Observe that the assigned values for these three parameter differ by orders of magnitude and represent the assumption that the three types differed in their abilities to tolerate high population density. By assumption, type 3 had the highest tolerance of high population density. In an attempt to assure the survival of the population under slow growth, it was assumed that the initial population contained only 10,000 individuals of type 1 and that no mutant types were present.

Figure 10.3.1 contains the graphs of the trajectories of the three types calculated for 6,000 generations of evolution by using the embedded deterministic model. From these trajectories, it can be seen that type 3, the type that was most tolerant of high population density, did not become predominate in the population until about 2,500 generations into the projection and did not reach a plateau of about 7×10^6 individuals until sometime after 3,000 generations of evolution. This rather slow rate of evolution was attributable to relatively small values of the λ-parameters, see above discussion.

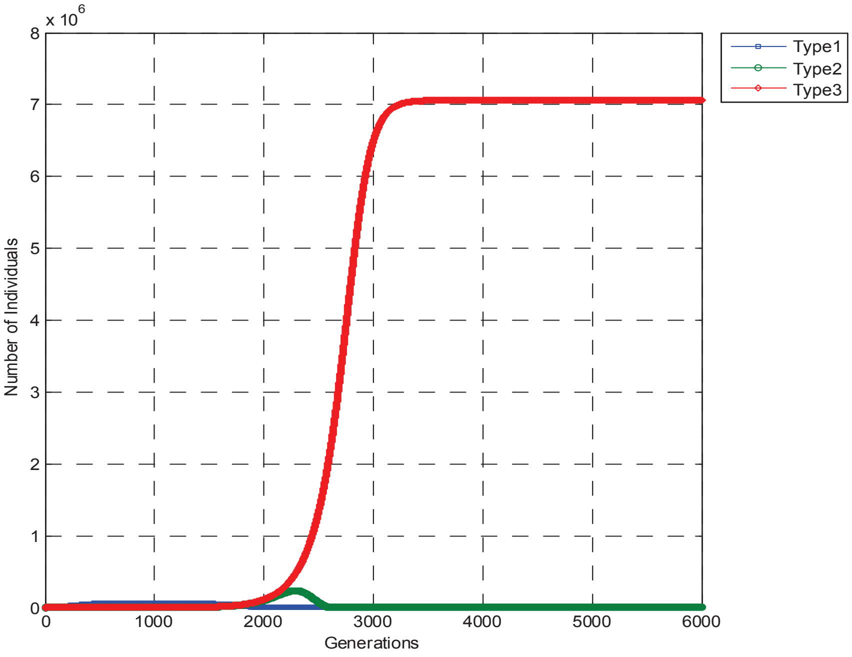

Figure 10.3.1 Evolution of the Population According to the Deterministic Model for Experiment 10.3.1.

Presented in Figure 10.3.2 are graphs of the estimates of the Q50 trajectories for each of the three types, using 100 replications of the simulated Monte Carlo data covering 6,000 generations of evolution. As can be seen from an inspection of the graphs of these trajectories, the pace of evolution for the stochastic model was more rapid than that portrayed by embedded deterministic model in figure 10.3.1. For from this figure it can be seen from the graph of the Q50 trajectory for type 3 that this type became predominant in the population a little after the midpoint of 1,000 and 2,000 generations; whereas, in the deterministic projection displayed in figure 10.3.1, the predominance of type 3 in the population occurred about midway between generations 2,000 and 3,000. For both types of projections, however, the trajectory of type 3 leveled off at about 7 × 10^6 individuals, while those for type 1 and 2 were still present in the population but in much smaller numbers.

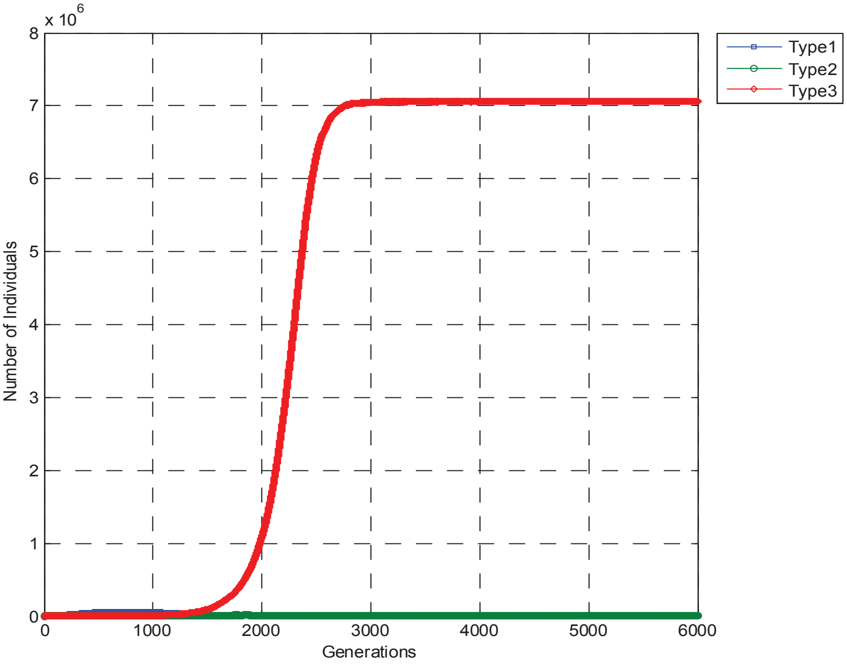

Figure 10.3.2 Evolution of Q50 Trajectories According to the Stochastic Model for Experiment 10.3.1.

In experiment 10.3.2, all parameter values used in experiment 10.3.1 were retained with the exception of the initial numbers of the three types. For this experiment, the initial number of individuals of type 1 was assigned the value 20 and it was assumed that mutant types 2 and 3 were not present in the population. The primary objectives of this experiment were to study the survival of the population and, if it did survive, what type of individual would become predominant in the population.

Displayed in Figure 10.3.3 are the graphs of the trajectories for the three types as computed using the embedded deterministic model. As expected, due to the small initial number of individuals of type 1, the pace of evolution of the population was much slower that in figure 10.3.1. For in this figure, type 3 reached its plateau of about 7×10^6 individuals sometime after 4,000 generations into the projection, while type 1 and 2 were still present in the population but in much smaller numbers.

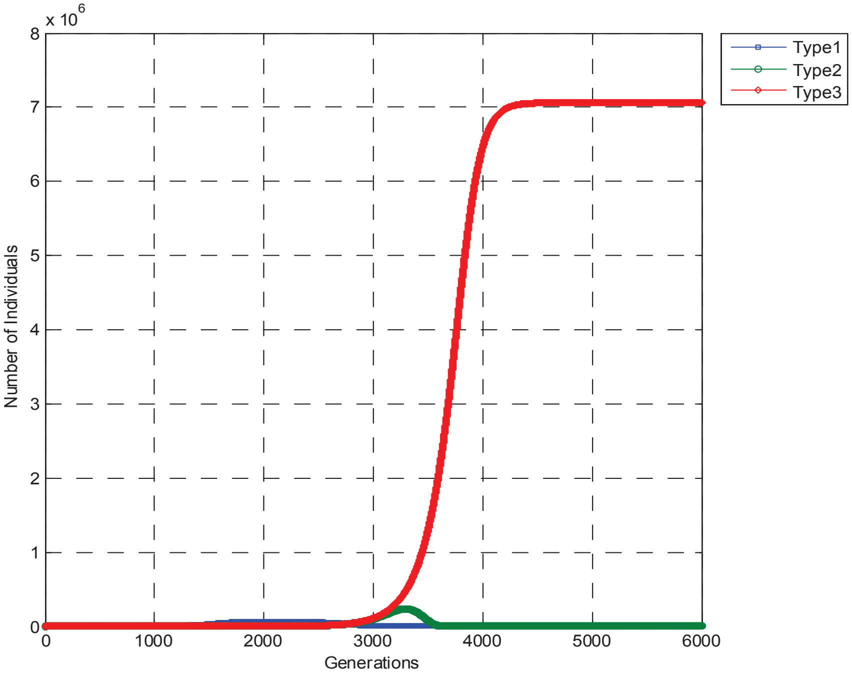

Figure 10.3.3 Evolution of the Population According to the Deterministic Model for Experiment 10.3.2.

By way of contrast, the graphs of the estimated MAX trajectories based on the Monte Carlo simulation data are displayed in Figure 10.3.4. Like those in figure 10.2.2, these graphs are based on the estimated 1 to 5 realizations among the 100 realizations of the process that did not become extinct in 6,000 or fewer generations. From an inspection of the MAX trajectory for type 3, it can be seen that the pace of evolution, for those realizations of the process in which they did not become extinct, was much faster than that estimated by using the embedded deterministic model. For in the stochastic projection, the MAX trajectory for type 3 reached a plateau of about 7×10^6 individuals at about 2,000 generations into the projection in contrast to about 4,000 generations for the deterministic projection in Figure 10.3.1. A serious limitation of the embedded deterministic model in experiment 10.3.2 is that allowance is not made in the formulation for the populations to become extinct as it is in a self regulating multitype branching process. Therefore, if an investigator were to rely solely on the deterministic model to project the evolution of a population, those cases in which the population becomes extinct would not be recognized.

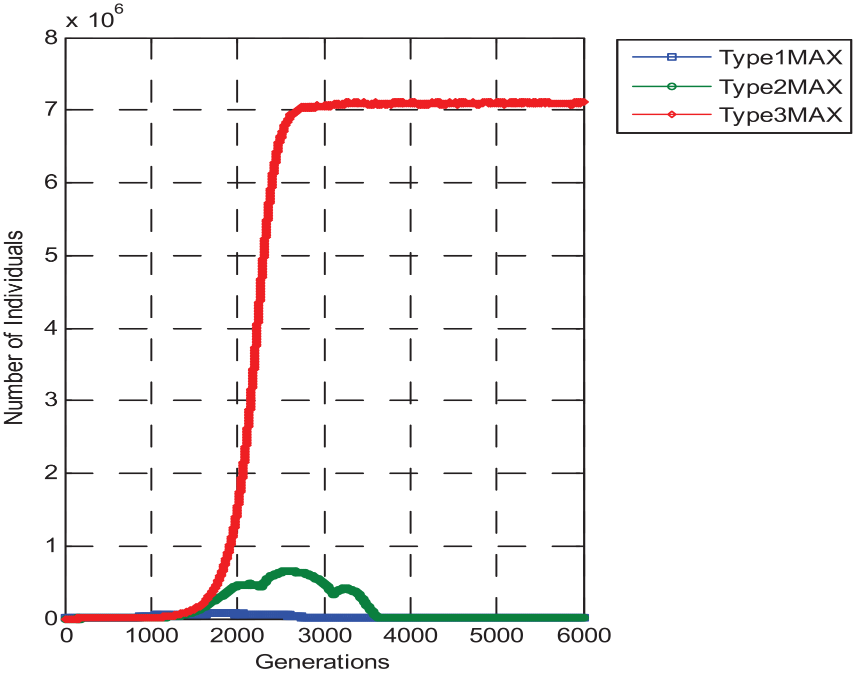

Figure 10.3.4 Evolution of MAX Trajectories According to the Stochastic Model for Experiment 10.3.2.

10.4 Survival of Deleterious and Beneficial Mutations From a Small Founder Populations

In experiment 10.4.1, the self regulating branching process with three types under consideration was applied to study the survival of deleterious and beneficial mutations when the founder population of three types consisted of the initial vector $\mathbf{Y}_0 = (1,000, 0, 0)$. Just as in experiments 10.3.1 and 10.3.2, the number of Monte Carlo replications of the experiment was 100 and for each replication the population was projected for 6,000 generations of evolution. Among the objectives of this experiment was to find whether the number of 1,000 initial individuals of type 1 was sufficient for occurrence and survival of a rare but beneficial mutation. The mutation probabilities were assigned the following values. It was assumed that type 1 may mutate to either type 2 or type 3 with probabilities $\mu_{12} = 10^{-5}$ and $\mu_{13} = 10^{-8}$ per generation. Observe that, in this experiment, type 3 was designated as the rare mutant. It was also assumed that type 2 could not back mutate to type 1 so that $\mu_{21} = 0$, but that a mutation to type 3 was possible with probability $\mu_{23} = 10^{-8}$. Lastly, it was assumed that there was no back mutation of type 3 to either types 1 or 2. At this point, it is useful to observe that the 3×3 matrix of mutation probabilities had the reducible structure

$$\mathfrak{M} = \begin{pmatrix} \mu_{11} & \mu_{12} & \mu_{13} \\ 0 & \mu_{22} & \mu_{23} \\ 0 & 0 & 1 \end{pmatrix}. \tag{10.4.1}$$

The idea that types 2 and 3 were, respectively, deleterious and beneficial mutations was characterized in the assignments of λ-parameter values for each type, which were chosen as $\lambda_1 = 1.05, \lambda_2 = 1.005$ and $\lambda_3 = 1.1$. According to these parameter assignments, mutants of type 2 were deleterious in the sense that on average each individual of this type would contribute only $\lambda_2 = 1.005$ offspring to the next generation; whereas, type 3 is a beneficial mutation in that on average individuals of this type would contribute $\lambda_3 = 1.1$ offspring to the next generation. Another way of viewing the deleterious and beneficial assignments of the parameters values is to observe that the three λ-parameters have the ordering $\lambda_2 < \lambda_1 < \lambda_3$. By way of completing the parameter assignments, the α and β parameters in the Weibull survival function were assigned the values $\alpha_i = 2$ and $\beta_i = 10^{-8}$ for $i = 1, 2, 3$.

Figure 10.4.1 contains the trajectories of the three types according to the embedded deterministic model for 6,000 generations of evolution. From this figure it can be seen that the number of individuals of type 1 rises to over 2×10^7 at about 200 generations into the projection and then starts to decline to a value near 0 by about generation 500. The number of deleterious type 2 mutants, however, remain near 0 throughout the projection. On the other hand, the number type 3 beneficial mutants in the population increases steadily from a low of zero in the initial generation a noticeable number at about 200 generations and then rises sharply to a plateau of over 3×10^7 individuals by generation of about 500.

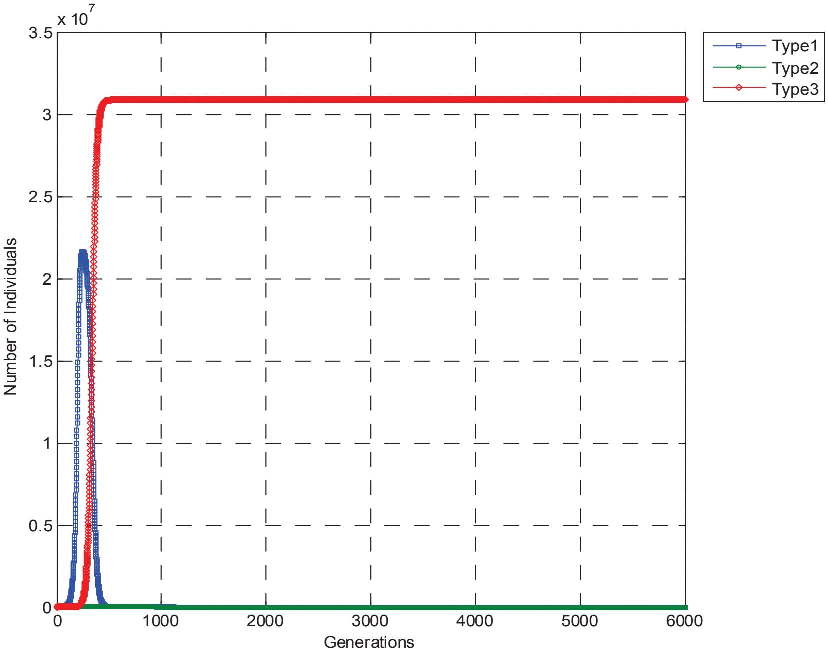

Figure 10.4.1 Trajectories of the Three Types for Experiment 10.4.1 According to the Deterministic Model.

Presented in Figure 10.4.2 are graphs of the MAX trajectories of types 1 and 2 for the stochastic model of experiment 10.4.1. As can be seen from this figure, the trajectory of type 1 rises from 1,000 individuals in the initial generation to over 2×10^7 individuals by about generation 500. Thereafter, it declines sharply to a values near to 0 at about 750 generations into the projection. The trajectory of deleterious type 2 mutant is either close to zero or zero throughout the projection. From an inspection of the simulated data, it was apparent that types 1 and 2 had become extinct by generation 1,000 into the projection

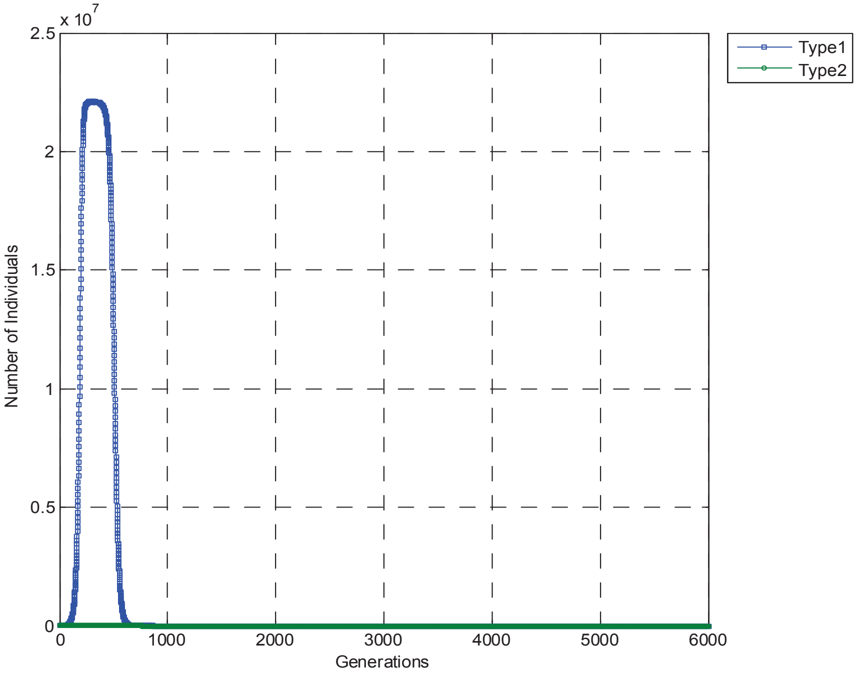

Figure 10.4.2 Max Trajectories in the Stochastic Model for Types 1 and 2 in Experiment 10.4.1.

Displayed in Figure 10.4.3 are the $MAX, MEAN, MIN$ and $Q50$ trajectories of beneficial mutant type 3 for Experiment 10.4.1. For this mutant type, all these trajectories had risen from an initial value of 0 to values near 3×10^7 at about 500 generations into the projection of 6,000 generations. After reaching a plateau of about 3×10^7 individuals, the trajectories remained close to each other in relative terms as can be seen from an inspection of the graphs after 1,000 generations. For experiment 10.4.1, in none of 100 Monte Carlo replications of the experiment did the individuals of beneficial mutant type 3 become extinct.

In experiment 10.4.2 all the parameter values used in experiment 10.4.1 were the same except that the initial number of individuals of type 1 was

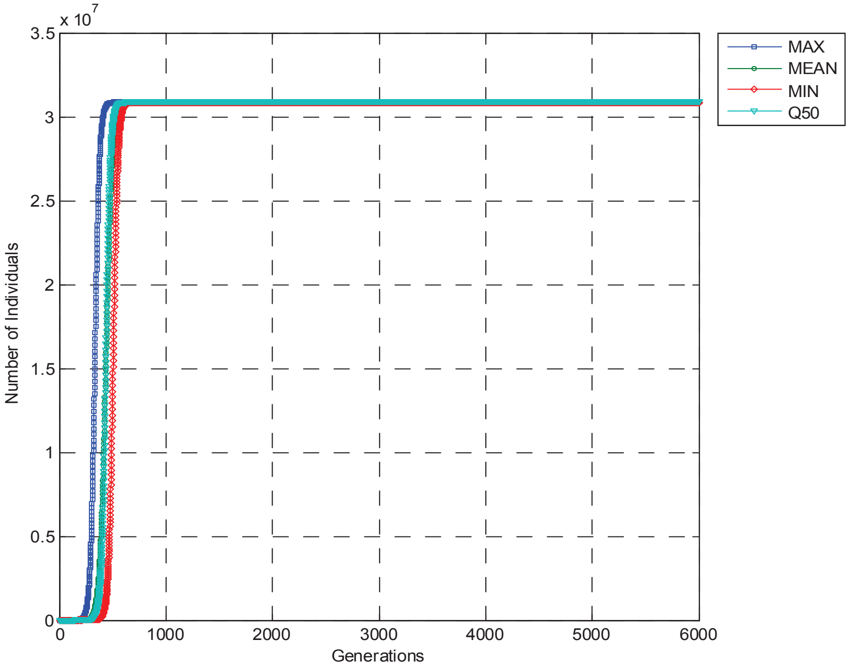

Figure 10.4.3 MAX, MEAN, MIN and Q50 Trajectories of Type 3 for Stochastic Model in Experiment 10.4.1.

20. The principal objective of this experiment was to test whether an initial founder population of this size was sufficient for beneficial mutant type 3 to become established in the population within 6,000 generations of evolution. For this experiment the graphs of the deterministic trajectories and those for the MAX for types 1 and 2 were similar to those in Figures 10.4.1 and 10.4.2, and will, therefore, be omitted. Presented in Figure 10.4.4 are the graphs of the $MAX, MEAN, Q25$ and $Q50$ trajectories for type 3 individuals in experiment 10.4.2.

As can be seen from this figure, the $MAX, Q25$ and $Q50$ trajectories for type 3 rise initially from 0 to a high of about 3×10^7 individuals during the span of 200 to 700 generations into the projection, but the $MEAN$ trajectory reached a plateau of only about 2×10^7 individuals at about 700 generations. The reason the $MEAN$ trajectory lies below that of the other trajectories was that, in experiment 10.4.2 for some of the replications of

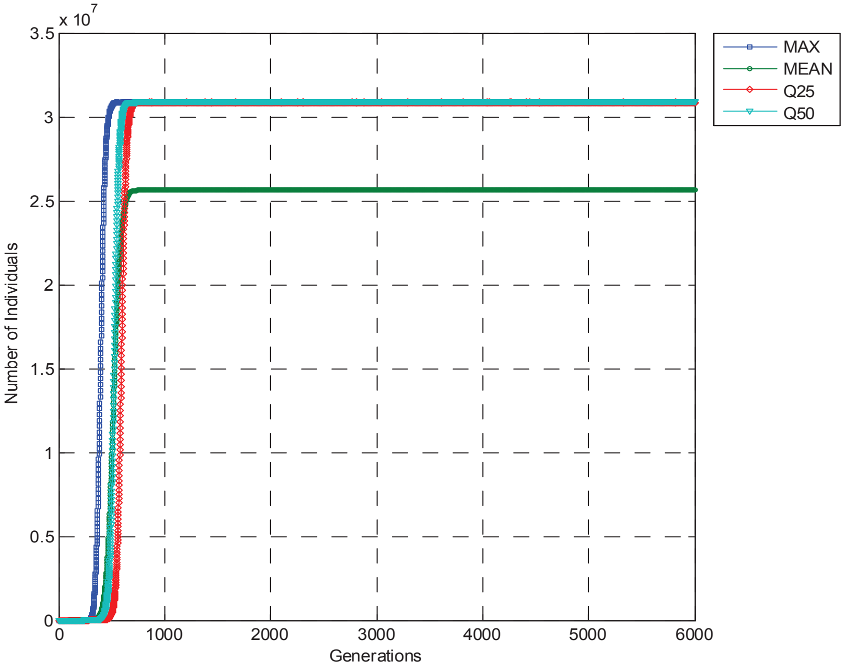

Figure 10.4.4 MAX, MEAN, Q25 and Q50 Trajectories of Type 3 for Experiment 10.4.2.

the process, type 3 individuals became extinct. In order to determine the frequency of replications for which extinction occurred for each of the three types, the software was modified to estimate these frequencies at generation 6,000. For the case of experiment 10.4.2, the estimates of these frequencies were, 1, 1 and 0.17 for types 1,2 and 3, respectively. Thus, for experiment 10.4.2, the estimated probability that individuals of beneficial type 3 survives and become predominant in the population was 0.83; whereas each of the types 1 and 2 became extinct with probability 1.

10.5 Survival of Mutations with Competitive Advantages Over an Ancestral Type

Experiment 10.5.1 was designed to study two components of natural selection; namely, the reproductive capacity of each type and the ability of each type to compete with the others for resources. With regard to reproductive capacity of the three types in a multitype self regulating branching process, it was assumed that the three λ-parameters, determining the expected number of offspring each type contributed to the next generation, were equal and had common value $\lambda_i = 1.05$ for types $i = 1, 2, 3$. Thus, the reproductive capacity component of natural selection in this experiment was neutral among the three types. This experiment differed from those described in section 10.3 in that the three types had the same reproductive capacity.

The abilities of the three types to compete for resources, however, were assumed to be different. More specifically, the β-parameters in the Weibull survival function were assigned the values $\beta_1 = 10^{-6}, \beta_2 = 10^{-7}$ and $\beta_3 = 10^{-8}$. Observe that the smaller the value of the β-parameter for a given type, the greater is the capacity of that type to compete for resources. It will be noted that in this experiment mutant types 2 and 3 were assumed to be more competitive than the ancestral type 1.

In particular, type 3 was assumed to be more competitive than type 2. Just as in experiments 10.4.1 and 10.4.2, the α-parameters in the Weibull survival function were assumed to have the common value $\alpha_i = 2$ for $i = 1, 2, 3$ and the initial vector for the population was assigned the value $\boldsymbol{Y}_0 = (1000, 0, 0)$, indicating that all individuals in the founder population were of the ancestral type 1. Lastly, the probabilities for mutations among the three types were assigned the same values as those in experiment 10.4.1 so that the mutation matrix had the reducible form displayed in (10.4.1).

Presented in Figure 10.5.1 are the graphs of the trajectories for the three types as estimated from the deterministic model embedded in the stochastic process for the first 2,000 generations of a projection of 6,000 generations, where the vertical axis is on a scale from 0 to 2.5×10^7. On this scale, the 1,000 initial individuals of type 1 are hardly noticeable but the trajectory for mutant type 2 had reached a value of about 0.23×10^7 or about $2,300,000$ individuals at about 300 generations into the projection and then declined along with type 1 to values close to zero for the rest of the projection. By about generation 500 of the projection, however, the trajectory for type 3, the type that was the most competitive among the three types, has a value of about 2.3×10^7 or about $23,000,000$ individuals and remained at this limit for the rest of the projection as shown in the figure.

Figure 10.5.2 contains the graphs of the MAX trajectories for types 1 and 2 for the first 2,000 generations of a projection by the stochastic process

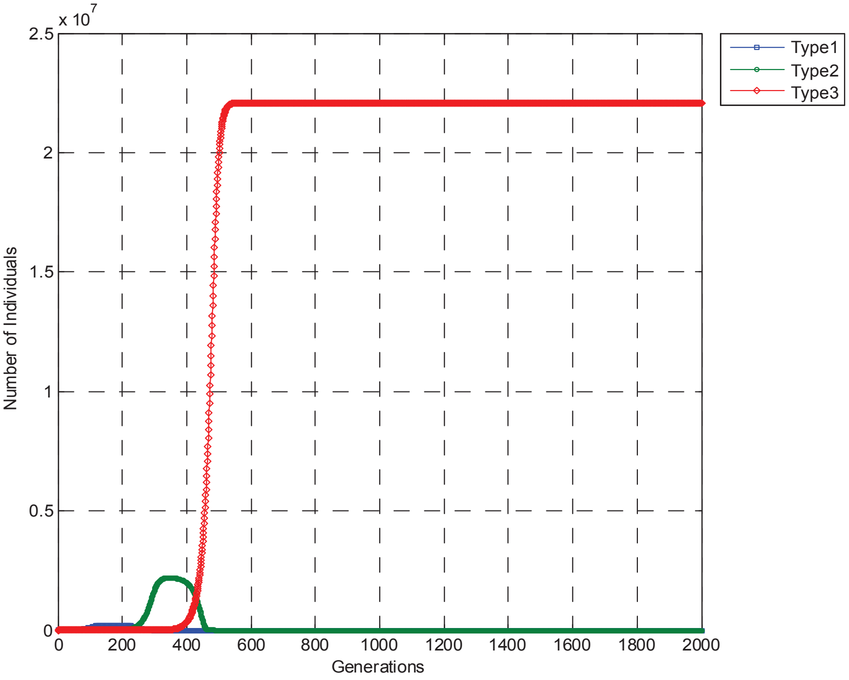

Figure 10.5.1 Trajectories of the Three Types for Experiment 10.5.1 According to the Deterministic Model.

for 6,000 generations, which were estimated from a sample of 100 Monte Carlo realizations of the process. In this figure, the vertical axis is on a scale from 0 to 2.5×10^6 and it can be seen that 1,000 initial individuals of type 1 reached a plateau of about 0.25×10^6 or $250,000$ individuals at about 100 generations into the projection but then declined to a number near 0 by generation 300. On the other hand, mutant type 2, which had a competitive advantage over the ancestral type 1, increased from 0 in the initial generation to a value of about 2.3×10^6 or about $2,300,000$ individuals by generation 300, but then declined to a values near 0 by about 450 generations of the projection.

As suggested by the graphs of the MAX trajectories in figure 10.5.2, types 1 and 2 became extinct sometime after about 600 generations into the projection. For this experiment, the software was modified to compute the frequency of extinction for each type at 6,000 generations into the projection. For experiment 10.5.1, the estimated frequency of extinction for both types 1 and 2 was 1, indicating that in all 100 realizations of the process types 1 and 2 had become extinct by 6,000 generations.

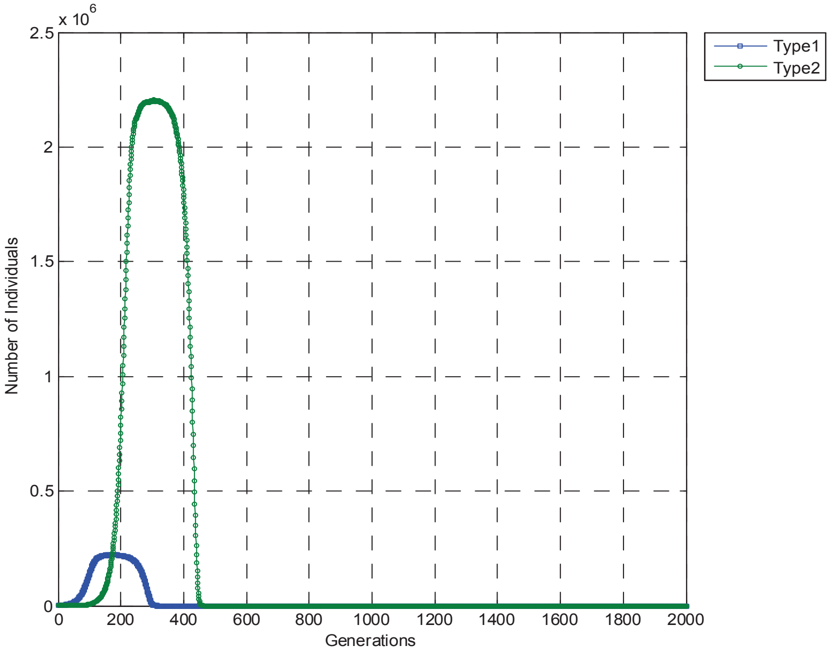

Figure 10.5.2 Max Trajectories in the Stochastic Model for Types 1 and 2 in Experiment 10.5.1.

Presented in Figure 10.5.3 are graphs of the $MIN, Q50, MAX$ and $MEAN$ trajectories for type 3 for the first 2,000 generations of a projection of 6,000 generations, where the vertical axis is on a scale from 0 to 2.5 × 10^7 Like the trajectories in figure 10.5.2, these trajectories were estimated from a sample of 100 Monte Carlo realizations of the process. As can be seen form this figure, the MAX trajectory for type 3 begins to rise from an initial values of 0 to a noticeable positive values at about 250 generations into the projection and then climbs rapidly to a value of about 2.3×10^7 or $23,000,000$ individuals at about 350 generations into the projection. On the other hand, the MIN trajectory of the process does not begin to rise to significant numbers until sometime after 400 generations into the projection.

This sharp difference between the times of increase for the MIN and MAX trajectories in the projection is indicative of the rather high levels of stochasticity that were present among the numbers of individuals of type 3 during first 400 generations of the projection. This high level of stochasticity is typical of the kind of erratic evolution that one would observe during the early stages of a process governed by the occurrence of rare mutational events.

Shortly after 400 generations into the projection, however, the process seem to converge to a statistical equilibrium in all trajectories are very close on the vertical scale of the graphs. This type of statistical equilibrium is typical for the stochastic process when the limit attained by the embedded deterministic model is a point of attraction.

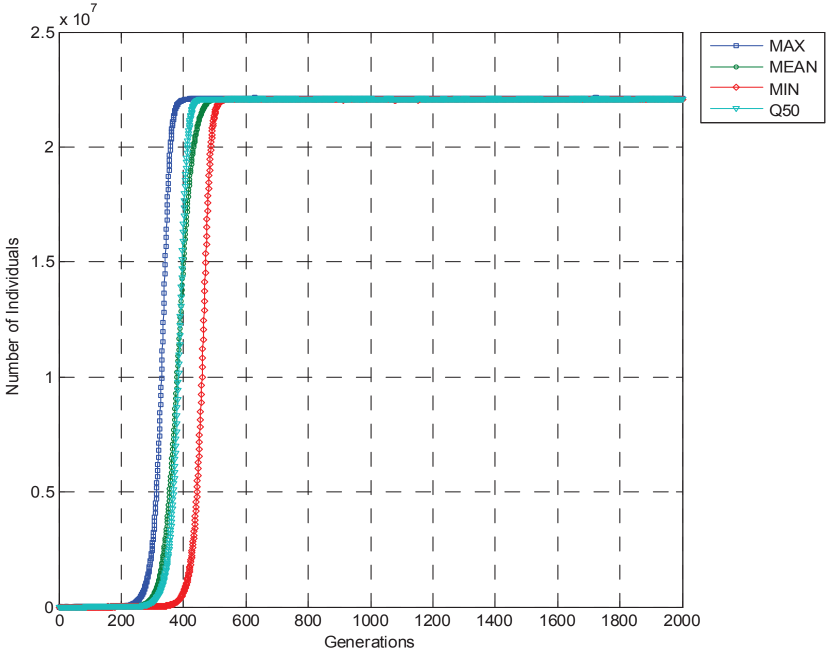

Figure 10.5.3 $MAX, MEAN, Q25$ and $Q50$ Trajectories of Type 3 for Stochastic Model in Experiment 10.5.1.

In experiment 10.5.2, all the parameters of the model were chosen as the same values as those in experiment 10.5.1 except that the initial vector for the population was chosen as $\boldsymbol{Y}_0 = (20, 0, 0)$, indicating that initial founder population consisted of only 20 individuals of type 1. The graphs for this experiment were quite similar to those for experiment 10.5.1 presented in Figures 10.5.1 and 10.5.2 above, and will, therefore, be omitted.

As expected, types 1 and 2 became extinct with probability one in this experiment. Presented in Figure 10.5.4 are graphs of the $Q25, Q50, MAX$ and $MEAN$ trajectories for experiment 10.5.2 for the first 2,000 generations of the projection. In this experiment, the frequency of extinction for type 3 at generation 6,000 was 0.17, indicating that at generation 6,000 in 17 of the 100 replications of the process type 3 individuals had become extinct. As an aside, it is interesting to note that in the 17 replications of the Monte Carlo simulation experiment, the entire population had become extinct.

Incidentally, this was the same result as that obtained in experiment 10.4.2, and there at least two explanations for this observation. Firstly, it could be merely a coincidence, or secondly, it could be due to the particular sets of random numbers used in these experiments.

One of the techniques used in all Monte Carlo simulation experiments reported so far in this book is that in each experiment, the same sequence of numbers is used as a seed to generate samples of random numbers. This is a very useful strategy in evaluating the results of Monte Carlo simulation experiments, because it is possible to repeat each experiment. On the other hand however, when two experiments give the same result, it seem likely that a factor underlying this outcome is that very similar sets of random numbers were used in both experiments.

As can be seen form this figure, the level of stochasticity among the 100 realizations during the interval of about 200 to 400 generations of the projection was very high, due to rare mutations and the small size of the initial population. Sometime after 600 generations into the projection the $MAX, Q50$ and $Q25$ trajectories are very close, indicating that among the realizations of the process for which extinction did not occur, the process converges to a statistical equilibrium. Interestingly, the $MEAN$ trajectory lies below that of the others after 600 generations, because by this point in the projection a significant number of realizations of the process for type 3 individuals had become extinct.

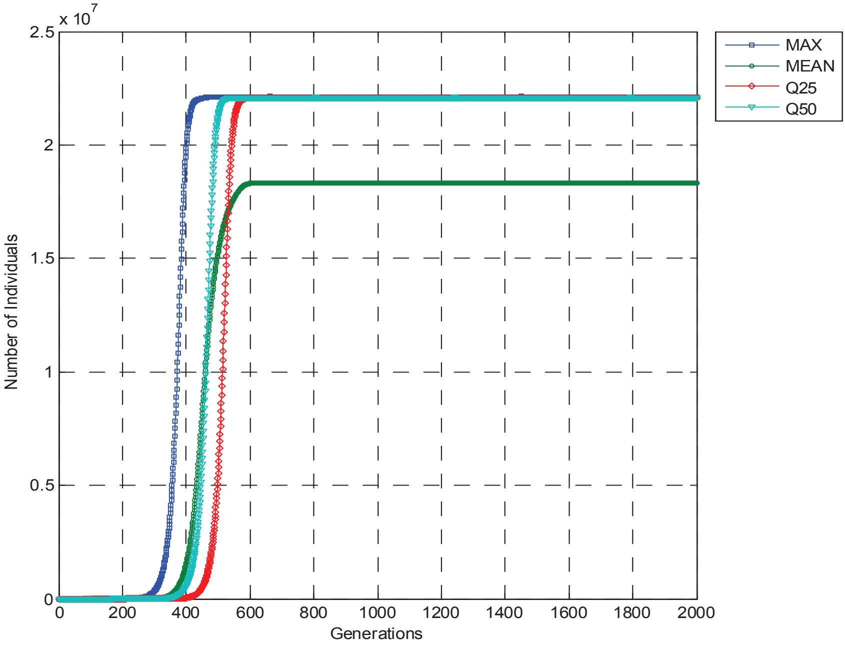

Figure 10.5.4 $MAX, MEAN, Q25$ and $Q50$ Trajectories of Type 3 for Experiment 10.5.2.

10.6 Chaotic Embedded Deterministic Model with Three Types

Among the goals of experiment 10.6.1 was to observe the behavior of the stochastic model when the trajectories of the embedded deterministic model exhibited chaotic behavior. As in all experiments considered in this chapter, the number of Monte Carlo replications of a projection of 6,000 generations was 100. Chaotic behavior of the embedded deterministic model is most often observed when the λ-parameters for the three types are quite large in comparison to the values used in other experiments reported in this chapter. In experiment 10.6.1, the λ-parameters had the value $\lambda_i = 5$ for $i = 1, 2, 3$ so that, by assumption, there was no selection with respect to reproductive capacity. It was also assumed that the mutation matrix was not reducible and had the numerical form

$$\mathfrak{M} = \begin{pmatrix} \mu_{11} & 10^{-5} & 10^{-8} \\ 10^{-5} & \mu_{22} & 10^{-8} \\ 10^{-8} & 10^{-9} & \mu_{33} \end{pmatrix}, \tag{10.6.1}$$

where the elements of the principal diagonal are chosen such that each row of the matrix sums to one.

As in all experiments reported in this chapter, the α-parameters in the Weibull survival function had the common values $\alpha_i = 2$ for $i = 1, 2, 3$ and the β-parameters had the values $\beta_1 = 10^{-6}, \beta_2 = 10^{-6.5}$ and $\beta_3 = 10^{-6.75}$. Observe that with parameter assignment, all the β-parameters had values in the interval $\left[10^{-7}, 10^{-6}\right]$ so that the range of β-values is smaller than those considered in previous sections. The motivation underlying this choice of values was to test whether the competitive advantages of types 2 and 3 would significantly change the behaviors the both the deterministic and stochastic trajectories of the model. Lastly, it was assumed that the initial vector had the form $\boldsymbol{Y}_0 = (500, 0, 0)$, indicating that the initial population consisted only of 500 individuals of type 1. The reason for considering this initial number, which is 0.5 of that considered in previous experiments, was the anticipation that population growth would be rapid, due to the relatively large values of the λ-parameters, and thus the likelihood of extinction was smaller than that considered in previous experiments.

Presented in Figure 10.6.1 are graphs of the trajectories for types 1 and 2 for the embedded deterministic model during the first 100 generations of the projection, where the vertical axis has the scale 0 to 6×10^6.

The reason for showing only the first 100 generations of the projection was that the rapid convergence of the trajectories to small values would have been obscured had more generations been included in the figure. From this graph it can be seen that both the trajectories for types 1 and 2 had decreased to values near 0 shortly after 20 generations into the projection. Prior to this time, the type 1 trajectory had reached a maximum of about 2×10^6 individuals at about 8 generations into the projection but had declined to near 0 at about 12 generations. On the other hand, the trajectory for type 2 had reached a much larger maximum of about 5.8×10^6 at about 18 generations before it declined, due to its superior competitive abilities of individuals of type 3.

Figure 10.6.2 contains the graph of the trajectory for individuals of type 3 for the first 100 generations of the projection as estimated from the embedded deterministic model. As can be observed from this figure, at about 25 generations into the projection the trajectory appears to settle

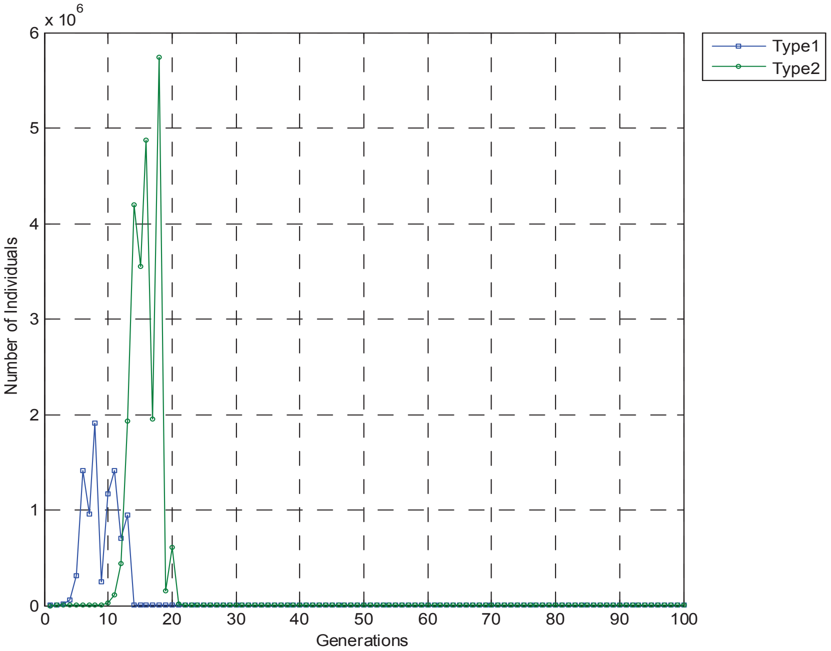

Figure 10.6.1 Graphs of the Deterministic Trajectories for Types 1 and 2 for the First 100 Generations in Experiment 10.6.1.

into a chaotic pattern in which population size ranges from a high of about 12×10^6 individuals to lows of about 1×10^6 individuals with other points ranging from about 2.5×10^6 individuals to an intermediate values of about 4.5×10^6 individuals. On the scale of 0 to 14×10^6 for the vertical axis, the fluctuations in the trajectory for type 3 individuals suggest that the deterministic model may have converged to a stable 10 cycle limit. When the simulated data were inspected on a one-generation time scale however, it was apparent that convergence to 10 constants had not occurred, even though the orders of magnitudes of the data points were similar to pattern displayed in figure 10.6.2.

As a check on whether the deterministic model had converged to a stable 10 cycle, the trajectories for type 3 were also inspected for generations 5,900 to 6,000. Presented in Figure 10.6.3 is the graph of the trajectory for type 3 for these generations. From this figure, it can be seen that the pattern shown in figure 10.6.2 persisted in the latter generations of a projection of

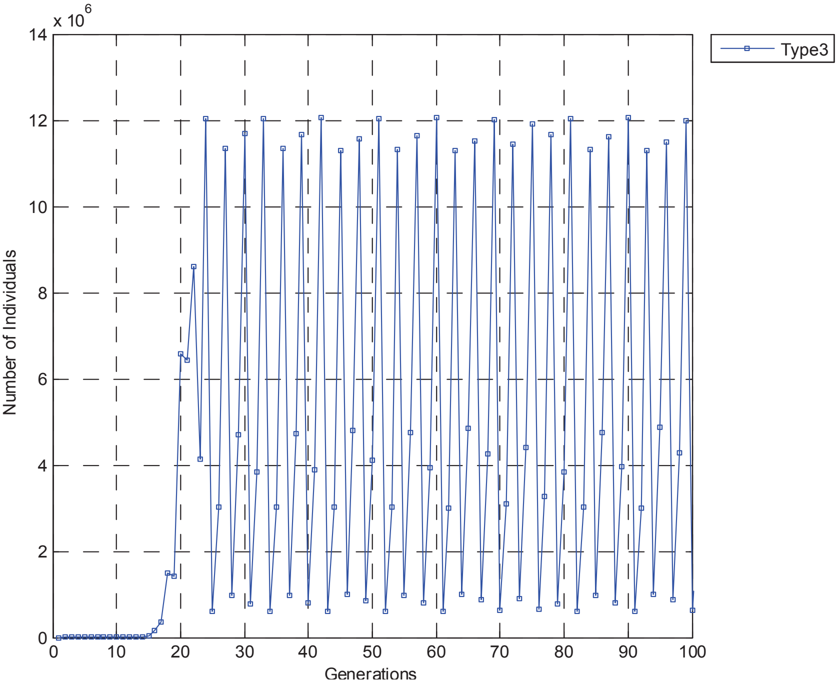

Figure 10.6.2 Graph of the Deterministic Trajectory for Type 3 for the First 100 Generations in Experiment 10.6.1.

6,000 generations. Furthermore, from an inspection of the simulated data for types 1, 2 and 3, it was clear that the deterministic model had not converged to a stable 10 cycle limit, which justifies using the term, chaotic, to describe the behavior of the trajectories. As expected, the trajectories of types 1 and 2 displayed fluctuations similar to those for type 3 but on much shorter scale for the vertical axis.

Displayed in Figure 10.6.4 are the graphs of the MAX and $MEAN$ trajectories for types 1 and 2 for the first 100 generations as estimated from a Monte Carlo simulation experiment with 100 replications. From this figure, it can be seen that the MAX and $MEAN$ trajectories had reached small values near zero on the graphs by about 15 generations into the stochastic projection. However, the corresponding trajectories for individuals of type 2 had not reached small values near zero until about 30 generations into the stochastic projection, which reflected the assumption that individuals of type 2 were more competitive than those of type 1. The superior

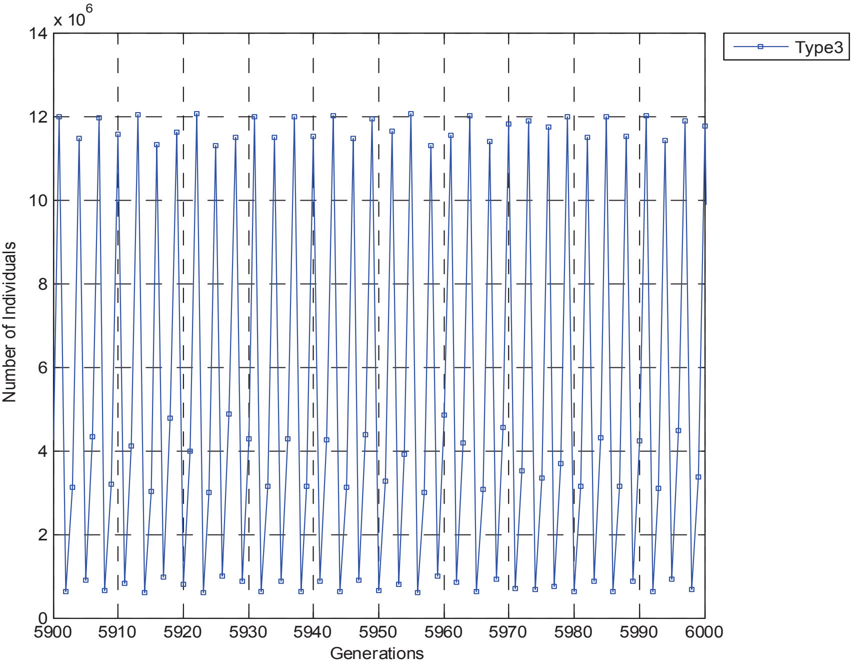

Figure 10.6.3 Graph of the Deterministic Trajectory for Type 3 for Generations 5,900 to 6,000 in Experiment 10.6.1.

competitive abilities of type 2 individuals was also apparent in the MAX trajectory for these individuals reached a high of nearly 7×10^6 individuals before it declined; whereas that for type 1 individuals reached a high of only about 2×10^6 individuals before it declined. That the MAX and $MEAN$ trajectories of both types 1 and 2 declined to small values was due to the assumption that type 3 was more competitive than both types 1 and 2.

To illustrate that types 1 and 2 were actually maintained in the population in small numbers, because of back mutations from type 3, the MAX and $MEAN$ trajectories for these types, which were estimated from the Monte Carlo simulation data for experiment 10.6.1, were also plotted for generations 5,900 to 6,000. Displayed in Figure 10.6.5 are graphs of these trajectories, where the vertical axis is on a scale of 0 to 30 individuals. From the graph of the MAX trajectory for type 2, it can be seen that the MAX trajectory is quite variable from generation to generation in that it varies from numbers in the neighborhood of 10 to values as high as 30 indi-

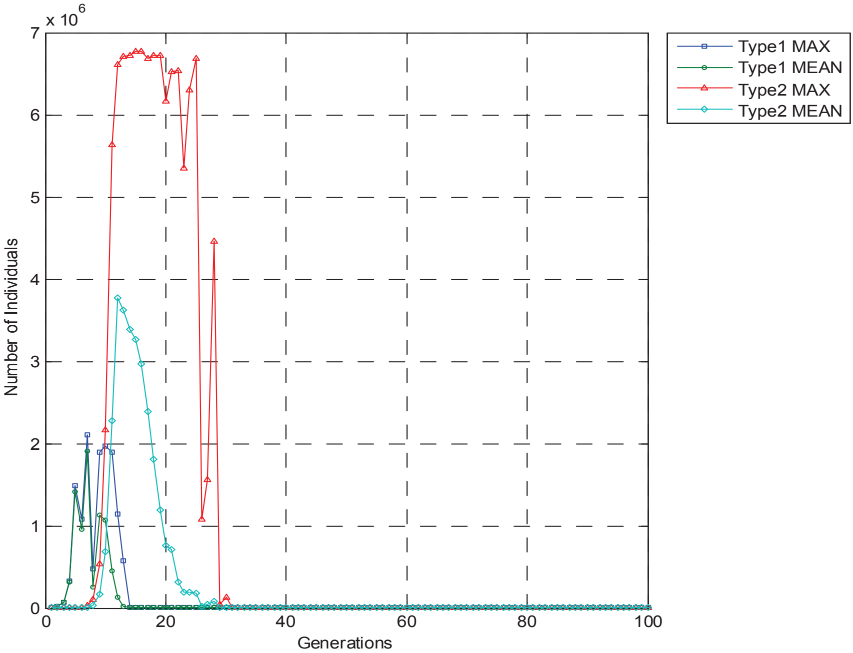

Figure 10.6.4 Graphs of the MAX and $MEAN$ Trajectories for Types 1 and 2 for the First 100 Generations in Experiment 10.6.1.

viduals. On the other hand, the MAX trajectory for type 1 is less variable from generation to generation in that it varies from values of a little less than 5 to values of 10 or slightly above. For both types 1 and 2, the MAX statistics represent outliers in the sense that the $MEAN$ trajectories for both types lie below the MAX trajectory for type 1. Another indicator of stochasticity due to small probabilities of back mutations from predominate type 3 was the estimates of the probabilities that neither individuals of type 1 nor type 2 were present in the population at generation 6,000 of the stochastic projection experiment. For the case of individuals of type 1, the estimate of this probability was 0.16; while that for type 2 was 0.04.

Presented in Figure 10.6.6 are the graphs of the $MAX, MEAN, MIN$ and $Q50$ trajectories for individuals of type 3 for the first 100 generations of the Monte Carlo projection for experiment 10.6.1. From this figure, it can be seen that the MAX trajectory begins to rise shortly after 10 generations into the projection and rises to a plateau of about 12×10^6 individuals by

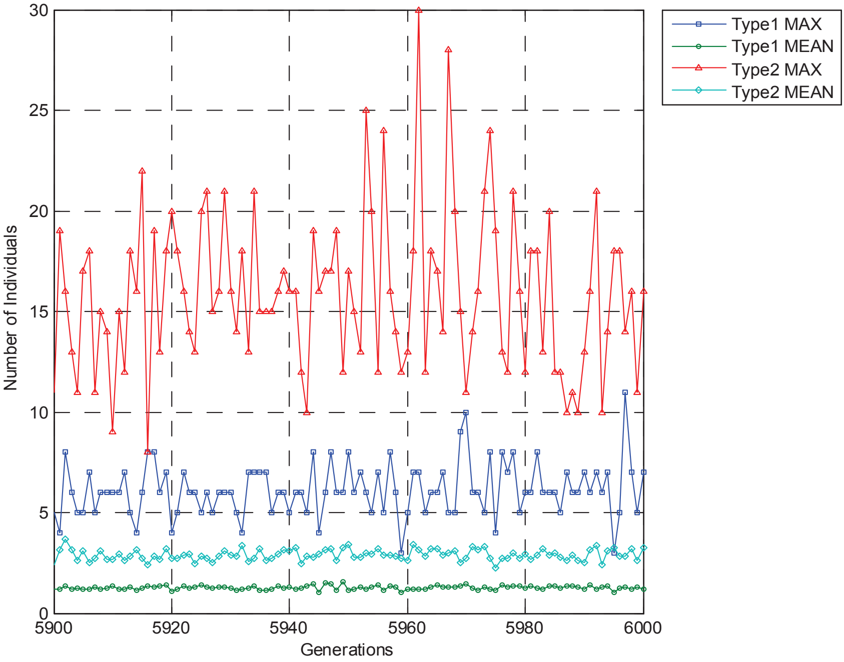

Figure 10.6.5 Graphs of the MAX and $MEAN$ Trajectories for Types 1 and 2 for Generations 5,900 to 6,000 in Experiment 10.6.1.

about generation 20. On the other hand, the MIN trajectory does not begin to rise until some time after generation 25 of the projection and rises to a plateau of less than 1×10^6 individuals in the neighborhood of 30 generations.

These differences in the times of increase for the MAX and $MEAN$ trajectories is characteristic of the high levels of stochasticity during the initial generations of stochastic projections when new mutations emerge from a more ancient parental type 1. After about generation 20, the $MEAN$ and the $Q50$ trajectories fall into a pattern of fluctuations in around a straight horizontal line.

For the case of the $MEAN$ trajectory this line appears to be somewhere between 5 and 6 million individuals; whereas the line for for $Q50$ trajectory lies somewhere between 3 and 4 million individuals. The pattern of the trajectories after 30 generations into the projection suggests that the stochastic population process had converged to a quasi-stationary distribution. As a further check on this idea, the $MAX, MEAN, MIN$ and $Q50$

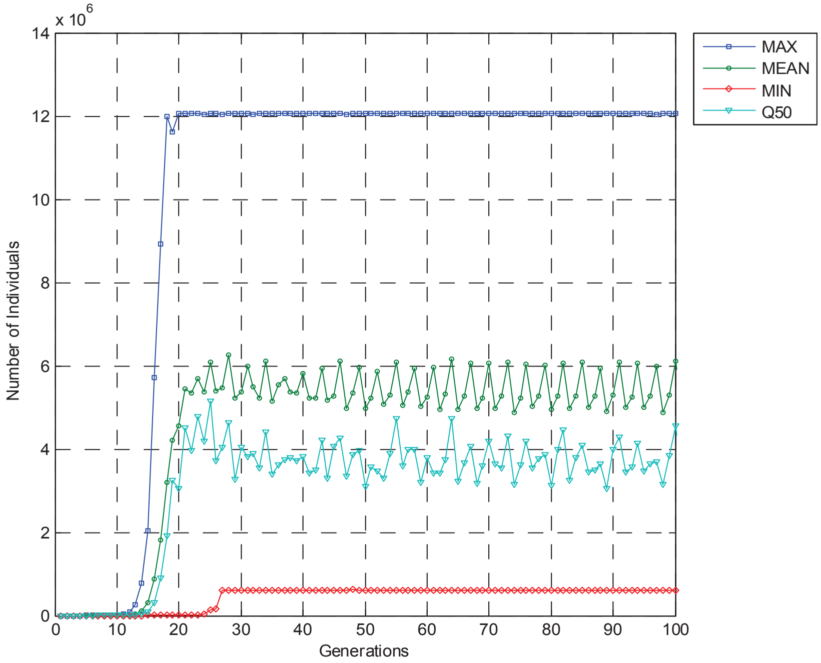

Figure 10.6.6 Graphs of the $MAX, MEAN, MIN$ and $Q50$ Trajectories for Type 3 for the First 100 Generations in Experiment 10.6.1.

trajectories were also plotted for generations 5,900 to 6,000 of the stochastic projection and are displayed in Figure 10.6.7. As can be seen from an inspection of this figure, it appears that the stochastic process had indeed converged to quasi-stationary distribution whose trajectories are similar to those observed in Figure 10.6.6.

In experiment 10.6.2, all the values of the parameters were the same as those in experiment 10.6.1 except that it was assumed that the founder population consisted of 20 individuals of type 1. Even in this experiment with a very small initial population, the probability that the population would become extinct was very low, due to the rather large expected number of offspring contributed to the next generation by each individual in a population. In this experiment, the estimate of the probability that no type 1 individuals were present in the population based on 100 replications in generation 6,000 was 0.12 and that for type 2 was 0.07. But, for type 3 individual, this estimate was 0, indicating that in generation 6,000 the

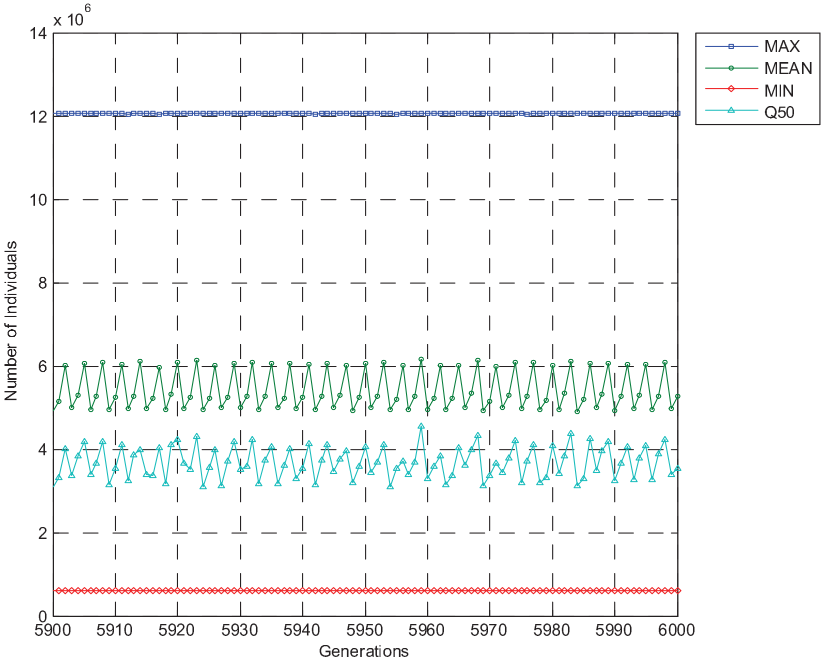

Figure 10.6.7 Graphs of the $MAX, MEAN, MIN$ and $Q50$ Trajectories for Type 3 for the Generations 5,900 to 6,000 in Experiment 10.6.1.

simulated data contained at least one individual of type 3 in each of the 100 replications. As the graphs of the kinds illustrated above were very similar to those for experiment 10.6.1, they will be omitted for experiment 10.6.2.

10.7 Self-Regulating Multitype Branching Processes in Random Environments

In this section, the self regulating branching process with three types studied in the preceding sections of this chapter will be generalized to include the effects of random environments that affect the values of vital parameters such as the expected number of offspring produced by each individual in any generation. Specifically, it will be assumed that environmental effects may be represented by first order autoregressive process (FOAR) introduced in chapter 7 that has the recursive form

$$X_i = \beta_0 X_{i-1} + \varepsilon_i \tag{10.7.1}$$

for generations $i = 1, 2, \ldots,$ where X_0 is a normally distributed random variable that must be computed in every realization of the process and $(\varepsilon_i \mid i = 1, 2, 3, \ldots)$ is a sequence of independent and identically distributed random variables with a common normal distribution with expectation 0 and variance σ^2. Furthermore, as was discussed in chapter 7, when the parameter β satisfies the condition $-1 < \beta_0 < 1$, the process is stationary and when it is in statistical equilibrium, each random variable has the variance

$$var\,[X_i] = \frac{\sigma^2}{1 - \beta_0^2} = \sigma_1^2 \tag{10.7.2}$$

for $i = 1, 2, \ldots.$ Therefore, to simulate realization of a FOAR process that is in a statistical equilibrium, it suffices to compute X_0 as a realization of a normal random variable with expectation 0 and with a variance σ_1^2 given (10.7.2) and then use the recursive equation (10.7.1) for $i = 1, 2, \ldots.$ Of course, if one wishes to simulate realizations of FOAR process in equilibrium with expectation μ, then one may compute the modified sequence $\mu + X_i$ for $i = 1, 2, \ldots.$

Just as shown in chapter 7, a finite sequence of realized values of the FOAR process $(x_i \mid i = 1, 2, \ldots, n)$ is mapped into the interval $(0, 1)$ by the mapping

$$u_i = \Phi\left(\frac{x_i - \mu}{\sigma_1}\right), \tag{10.7.3}$$

for $i = 1, 2, \ldots, n$ to yield a sequence of realized and correlated uniform random variables on the interval $(0, 1)$, where $\Phi(\cdot)$ is the standard normal distribution function. These realizations of correlated uniform random variables are in turn mapped into three intervals, representing the range of the three λ-parameters of a three-type self regulating branching process in a random environment. The three intervals for the λ-parameters will be denoted by

$$[\theta_{\nu 1}, \theta_{\nu 2}), \tag{10.7.4}$$

where $0 < \theta_{\nu 1} < \theta_{\nu 2}$ for type $\nu = 1, 2, 3$. Unlike the constant λ-parameters in the formulations of multitype branching process considered in chapter 9, in this formulation these parameters are realizations of random variables in each generation n given by the formula

$$\lambda_{n\nu} = \theta_{\nu 1} + (\theta_{\nu 2} - \theta_{\nu 1})\, u_{n\nu} \tag{10.7.5}$$

for each type $\nu = 1, 2, 3$, where $u_{n\nu}$ is a realization of the uniform random variable on the interval $(0, 1)$. When implementing this formula in a Monte Carlo simulation experiment with $G \geq 1$ generations in each replication, an array $n_1 = 3 \times G$ realizations of correlated uniform random variables on the interval $(0, 1)$ must be computed and arranged in a $G \times 3$ array in the computer. Each row of this array will have the form

$$\boldsymbol{u}\left(n\right) = \left(u_{n1}, u_{n2}, u_{n3}\right) \tag{10.7.6}$$

for each generation $n = 1, 2, \ldots, G$, which will be used in the computer implementation of formula $(10.7.5)$.

As the λ-parameters are random variables and are thus not constant from generation to generation, it is not possible to embed a deterministic model in the stochastic process following the procedure used in chapter 9, see section 9.9. The rationale underlying this statement is that, by assuming the λ-parameters are random variables, it can be shown that, by using the procedures discussed in chapter 7, that the self regulating multitype branching process under consideration is a mixture of Markov chains. Furthermore, the mixing process is derived from a Gaussian process based on FOAR model, and, within this mixing framework, the derivation of formulas for conditional expectations needs to be considered carefully. For example, within this framework, equation (9.9.10) takes the form

$$E\left[\boldsymbol{Y}\left(n+1\right) \mid \boldsymbol{Y}\left(n\right) = \boldsymbol{y}\left(n\right), \boldsymbol{u}\left(n\right)\right] = \boldsymbol{y}\left(n\right) \boldsymbol{\Lambda}\left(\boldsymbol{y}\left(n\right), \boldsymbol{u}\left(n\right)\right), \tag{10.7.7}$$

where the matrix $\boldsymbol{\Lambda}$ on the right is a function of both $\boldsymbol{y}\left(n\right)$ and $\boldsymbol{u}\left(n\right) = \left(u_{n1}, u_{n2}, u_{n3}\right)$. Explicitly, this matrix has the form,

$$\boldsymbol{\Lambda}\left(\boldsymbol{y}\left(n\right), \boldsymbol{u}\left(n\right)\right) = (\lambda_{n\nu} p_{\nu}\left(\boldsymbol{y}\left(n\right)\right) \mu_{\nu'\nu}), \tag{10.7.8}$$

see (9.9.9) and (10.7.5) for $n = 0, 1, 2, \ldots$. Briefly, the technical details will be left to the reader, but if one assigns the initial population vector $\boldsymbol{y}\left(0\right)$, then (10.7.8) for $n = 1$ becomes

$$\boldsymbol{\Lambda}\left(\boldsymbol{y}\left(0\right), \boldsymbol{u}\left(0\right)\right) = (\lambda_{0\nu} p_{\nu}\left(\boldsymbol{y}\left(0\right)\right) \mu_{\nu'\nu}), \tag{10.7.9}$$

where $\lambda_{0\nu}$ is calculated as in $(10.7.5)$, so that one could use this estimate to calculate $\widehat{\boldsymbol{Y}}\left(1\right)$ as

$$\widehat{\boldsymbol{Y}}\left(1\right) = \boldsymbol{y}\left(0\right) \boldsymbol{\Lambda}\left(\boldsymbol{y}\left(0\right), \boldsymbol{u}\left(0\right)\right). \tag{10.7.10}$$

It is then possible to proceed as in chapter 9 to calculate the sequence $\widehat{\boldsymbol{Y}}\left(n\right)$ recursively for $n = 1, 2, \ldots, G$ for each replication of the process, but the sequence so calculated would be stochastic in the sense that it would vary among replications or realizations of a process that evolves for

G generations. In this section, however, the ideas just described were not implemented so that all the experiments discussed in the remainder of this section will be based only the stochastic model.

In experiment 10.7.1, the number of replications of the experiment was chosen as 100, and, within each replication, 2,000 generations of evolution were computed. As many more realizations of uniform random variables on the interval $(0,1)$ needed to be computed in this experiment than for those discussed in the foregoing sections of this chapter, the computer time taken to complete the experiment was 24.5 hours on a computer with a speed of 3 giga hertz. The ranges of the theta parameters for the three types, see (10.7.4) were assigned the values $\theta_{11} = 0.95, \theta_{12} = 2, \theta_{21} = 0.95, \theta_{22} = 3, \theta_{31} = 0.95$ and $\theta_{32} = 4$. The rationale underlying these choices was the desire to run an experiment in which the lengths of the intervals in (10.7.4) were quite large so that the expected number of offspring produced by each individual may vary significantly from generation to generation. The 3×3 matrix of mutation probabilities $\mathfrak{M}$ was chosen as in (10.6.1), and the α and β parameters in the Weibull survival function for the three types of individuals were chosen as $\alpha_\nu = 2$ and $\beta_\nu = 10^{-6}$ for $\nu = 1, 2, 3$. With respect to the FOAR process in (10.7.1), β_0 was assigned the value $\beta_0 = 0.9$ so that neighbors in a realization of the process would be highly correlated, and standard deviation σ was assigned the value $\sigma = 2$, which yields a values of 21.0526 for the variance given by formula (10.7.2) with a standard deviation of $\sigma_1 = 4.5883$. Finally, the founder population for this experiment consisted of 1,000 individuals of type 1.

Presented in Figure 10.7.1 are the graphs of the $MIN, Q25, Q50, Q75$ and MAX trajectories for the first 500 generations of the experiment. As can be seen form the MAX trajectory in this figure, significant numbers of type 3 individuals did not appear in the population until about 20 to 25 generations into the projection, but by 50 generations, the MAX trajectory reached values close to 16×10^5 individuals.

Moreover, the relative flatness of all graphs of the trajectories in the figure suggests that the process had converged to a quasi-stationary distribution by about 100 generations into the projection in which individuals of type 3 were predominant in the population. Not shown in the figure was the presence of small numbers of individuals of types 1 and 2. To provide further evidence that the process had indeed converged to a quasi-stationary distribution at about 100 generations, the graphs of the trajectories under consideration were also presented in Figure 10.7.2 for generations 1,900 to 2,000. As can be seen from this figure, the graphs of these trajectories fluctuate around values close to those in Figure 10.7.1 at about 100

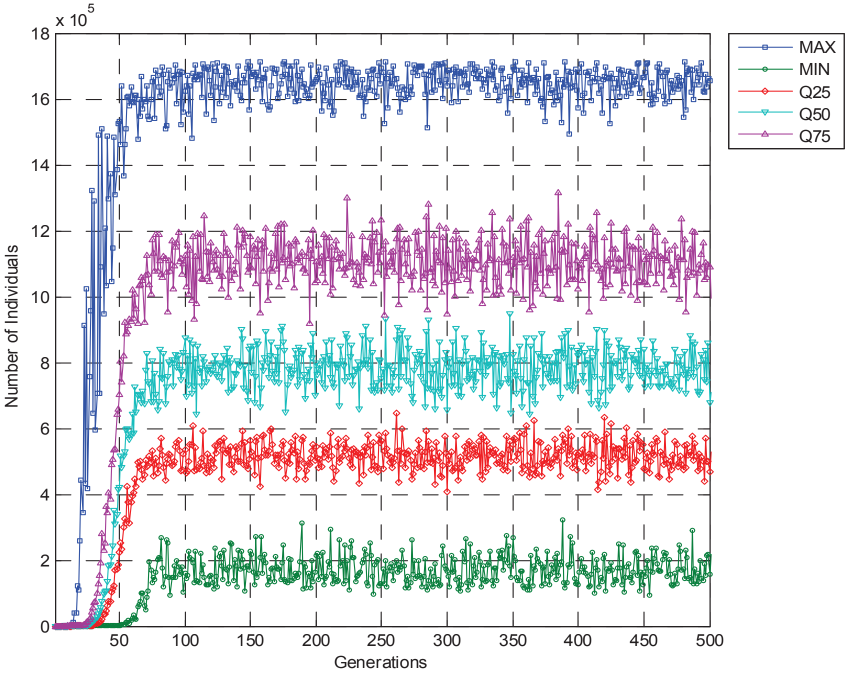

Figure 10.7.1 Graphs of the $MIN, Q25, Q50, Q75, MAX$ Trajectories for Type 3 Individuals in Experiment 10.7.1 for Generations 1 to 500.

generations, which provides evidence that a quasi-stationary distribution had been reached in about 100 generations. In populations characterized by the parameter assignments in this experiment, evolution to a statistical equilibrium would be quite rapid.

In experiment 10.7.2, all the parameter values chosen for experiment 10.7.1 were also chosen for this experiment except the ranges of the λ-parameters for the three types were restricted to intervals with smaller ranges and the number of generation considered was smaller. In particular, the θ-parameters for this experiment were assigned the values $\theta_{11} = 0.95, \theta_{12} = 1.2, \theta_{21} = 0.95, \theta_{22} = 1.3, \theta_{31} = 0.95$ and $\theta_{32} = 1.4$, which resulted in a much slower pace of evolution than that observed in experiment 10.7.1. To reduce the amount of computer time needed to complete an experiment with 100 replications, the evolution of the population was considered for only 500 generations for each replication. For this ex-

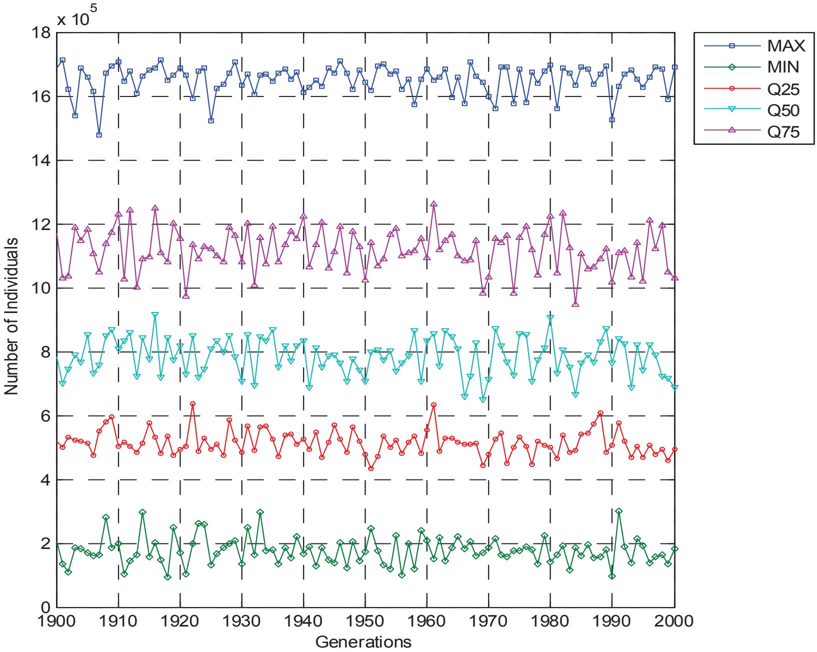

Figure 10.7.2 Graphs of the $MIN, Q25, Q50, Q75, MAX$ Trajectories for Type 3 Individuals in Experiment 10.7.1 for Generations 1,900 to 2,000.

periment, only about 1.5 hours of computer time were required to complete the simulation.

Figure 10.7.3 contains the graphs of the $MIN, Q25, Q50, Q75$ and MAX trajectories for this experiment for 500 generations. From this figure it can be seen that the pace of evolution for this population was much slower than that for experiment 10.7.1. For in this experiment, the MAX trajectory did not reach a significant number until somewhere between 75 and 100 generations into the projection. Furthermore, the graphs of the trajectories suggest that the process did not converge to a stationary distribution until the MIN trajectory reached values on the range to 2×10^5 to 3×10^5 somewhere between 350 and 400 generations into the projection.

In experiment 10.7.3, all the values of experiment 10.7.2 were retained except that the β_0 for the FOAR process was changed to $\beta_0 = -0.9$. With this assignment for the parameter β_0, it is of interest to note that $var\left[X_0\right]$ in

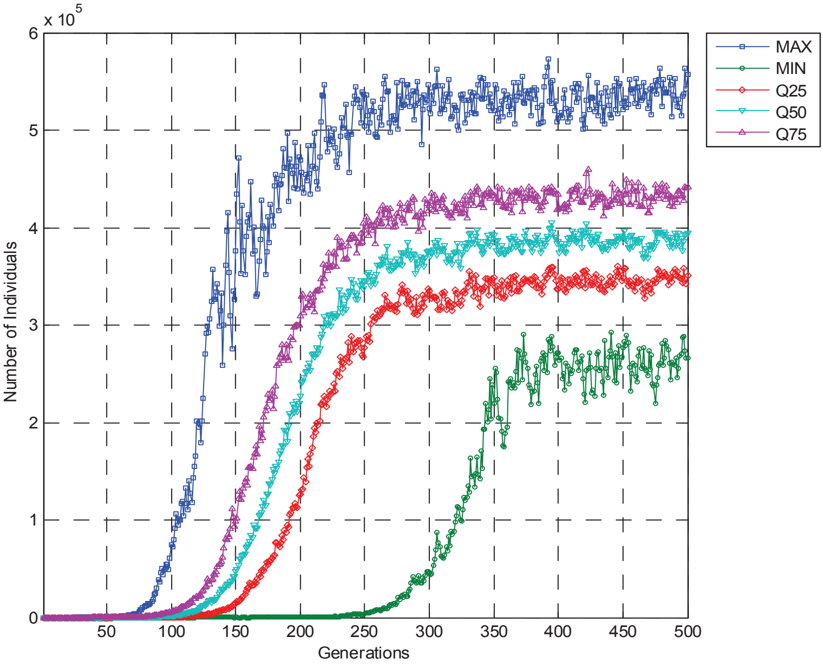

Figure 10.7.3 Graphs of the $MIN, Q25, Q50, Q75, MAX$ Trajectories for Type 3 Individuals in Experiment 10.7.2 for Generations 1 to 500.

the FOAR process has the same value as that in experiment 10.7.1. Shown in Figure 10.7.4 are the graphs of the trajectories of type 3 individuals for this experiment.

As can be seen from this figure the level of stochasticity among the realizations of this process was much greater than in other two experiments presented in this section, which was due to the negative value assigned to the parameter β. For in this experiment, the MIN trajectory did not rise to a significant number of individuals of type 3 until nearly 500 generations into the projection. Such a result indicates that among the 100 replications of the process significant numbers of type 3 individuals did not appear in the population until nearly 500 generations of evolution. It is of interest to note, however, that from the $Q50$ trajectory at 500 generations it can be seen that in 50% of the replications of the process, the number of individuals of type 3 in the population was greater than or equal to 3.5×10^5.

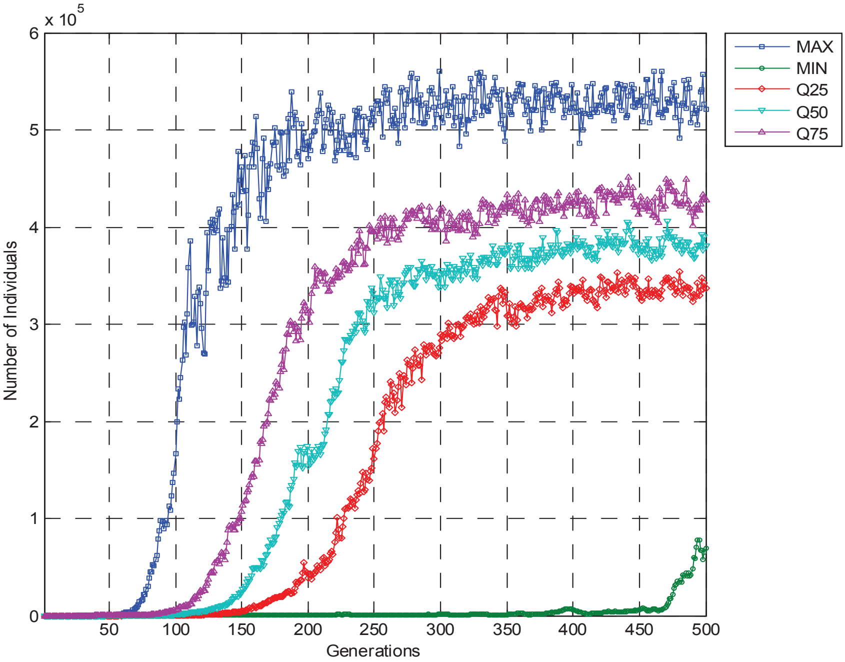

Figure 10.7.4 Graphs of the $MIN, Q25, Q50, Q75, MAX$ Trajectories for Type 3 Individuals in Experiment 10.7.3 for Generations 1 to 500.

10.8 Simulating Multitype Genealogies and Further Reading

In principle, given the outline of ideas on simulating multitype genealogies presented in section 9.7 and the examples of self regulating branching processes with three types described in this chapter, it would be possible to write software to simulate multitype genealogies for each of the parameters choices investigated in the foregoing sections of this chapter. But, the challenging task of writing of such software will be left as exercises for an interested reader, who may use a programming language of his choosing. In the remainder of this section, other literature on formulating and analyzing self regulating branching process will be discussed. Quite often the class of branching processes, named self regulating in this book, is referred to as population density dependent processes by other authors.

An example of an alternative approach to the formulation of the offspring distributions for the case of one type density dependent branching process is that described by G. Hognas in section 6.9 of the book by Haccou, Jagers and Vatutin (2007). Let the random variable N taking values in the set $(i \mid i = 0, 1, 2, 1 \ldots)$ of non-negative integers denote the number of offspring produced by each individual in any generation of a branching process, and let $\exp(-\beta i)$, where $\beta > 0$ is a parameter, denote the probability that an individual survives to reproduce, given a population size of i.

Given that an individual survives, let p_k denote the probability that an individual contributes k offspring to the next generation, where $p_k \geq 0$ for all $k \geq 0$ and

$$\sum_{k=0}^{\infty} p_k = 1. \tag{10.8.1}$$

Let the random variable Y_n denote the size of the population in generation n. Then, according the formulation described by Hognas,

$$P\left[N = k \mid Y_n = i\right] = \exp\left(-\beta i\right) p_k \tag{10.8.2}$$

for $k = 1, 2, \ldots$ so that

$$P\left[N \geq 1 \mid Y_n = i\right] = \exp\left(-\beta i\right) \sum_{k=1}^{\infty} p_k = \exp\left(-\beta i\right) \left(1 - p_0\right). \tag{10.8.3}$$

Therefore, according to this formulation,

$$P\left[N = 0 \mid Y_n = i\right] = 1 - \exp\left(-\beta i\right) \left(1 - p_0\right). \tag{10.8.4}$$

It is of interest to note that this formulation is different from that presented in section 9.4, where a description of the formulation of the offspring distribution of self regulating branching processes used in this book was presented. In particular, it will be helpful to look at the implications of formulas (9.4.1) and (9.4.2) in terms of generating functions. If, as in section 9.4, let $p\left(y_1\right)$ denote the conditional probability that an individual survives to reproduce, given that population size is y_1, and suppose the distribution of the random variable N has the general form described in equation (10.8.1).

Then, given that $N = n$, the conditional generating function of the sum on the right in (9.4.1) is

$$\left(p\left(y_1\right) s + 1 - p\left(y_1\right)\right)^n. \tag{10.8.5}$$

Let

$$G\left(s\right) = \sum_{k=0}^{\infty} p_k s^k \tag{10.8.6}$$

denote the generating function of the random variable N. Then, the generating function of the random variable $\varsigma\left(y_1\right)$, representing the number of offspring produced by any individual, in the left had side of (9.4.1) is

$$H\left(s\right)=G\left(\left(p\left(y_1\right)s+1-p\left(y_1\right)\right)\right)=\sum_{k=0}^{\infty}p_k\left(p\left(y_1\right)s+1-p\left(y_1\right)\right)^k, \tag{10.8.7}$$

which is defined for all s such that $0\leq s\leq 1$. Therefore, all offspring distributions used in this chapter and chapter 9 of this book are special cases of the class of distributions characterized by compound generating functions in (10.8.7). For this class of distribution,

$$\begin{aligned}P\left[\varsigma\left(y_1\right)=0\right]&=H\left(0\right)=G\left(1-p\left(y_1\right)\right)\\&=\sum_{n=0}^{\infty}p_n\left(1-p\left(y_1\right)\right)^n,\end{aligned} \tag{10.8.8}$$

which is clearly different form formulas (10.8.3) and (10.8.4). Unlike the formulation described by Hognas, equation (10.8.8) follows from the formulation set forth in section 9.4 and there is no need to force the extra definition given in formula (10.8.4).

As has been demonstrated in chapter 9 and in this chapter, the method used to formulate offspring distributions in one and multitype branching processes is effective in dealing with the issues motivating the computer experiments presented in these chapters in that it allows for variations in population size among generations and replications of realizations of a stochastic process. Other approaches to formulating density dependent processes are also discussed by Hognas, including the famous logistic model, which lead to the development of theories of chaos. The book by Haccou, Jagers and Vatutin (2007) may be consulted for an extensive list of references on branching processes and other processes of population dynamics. Interestingly, in his book Cohen (1995) on "How Many People Can the Earth Support?" also used a version of the logistic model what was termed mathematical cartoons of human population size and carrying capacity in a curiously humorous approach to serious issues confronting mankind.

Bibliography

[1] Cohen, J. E. (1995) **How Many People Can the Earth Support?** W.W. Norton & Company, New York and London.

[2] Fisher, R. A. (1958) **The Genetical Theory of Natural Selection - Second Revised Edition.** Dover Publications, New York.

[3] Haccou, P., Jagers, P. and Vatutin, V. A. (2007) **Branching Processes - Variation, Growth and Extinction of Populations**. Cambridge University Press, The Edinburgh Building, Cambridge, CB2 8Ru, UK.

[4] Hazen, R. M. (2005) **Origins of Life**. The Teaching Company. Chantilly, Virginia, USA. www.teach12.com.

[5] Pigliucci, M. and Kaplan, J. (2006) **Making Sense of Evolution - The Conceptual Foundations of Evolutionary Biology**. The University of Chicago Press, Chicago and London.

[6] Okasha, S. (2006) **Evolution and the Levels of Selection**. Oxford University Press, Oxford and New York.

[7] Sober, E. (2006) **Conceptual Issues in Evolutionary Biology, Third Edition**. The MIT Press, Cambridge, Mass. and London, England.

Chapter 11

Two Sex Multitype Self-Regulating Branching Processes in Evolutionary Genetics

11.1 Introduction

In this chapter, stochastic models for the evolution of a diploid population with two sexes are formulated and used in Monte Carlo simulation experiments designed to study the evolution of a population under various assumptions about mutation and three components of natural selection. Like the models considered in chapters 9 and 10, in this chapter attention will be focused on populations that evolve in discrete time generations. Briefly, the three components of selection are incorporated into the model in terms of three sub-modules. One module focuses on choices of sexual partners for each sex, which are characterized in terms of acceptance probabilities indicating whether a female or male with a given genotype or phenotype may prefer sexual partners with certain genotypes or phenotypes over others. This module of the formulation is sufficiently flexible to include random and various types of assortative mating by assigning different numerical values to the set of acceptance probabilities. A second module characterizes reproductive success for each mating type, consisting of various genotypic combinations of females and males, in terms of the expected number of offspring produced by each couple type. The third module of the formulation consists of survival probabilities of the Weibull type characterizing the survival of offspring produced in a given generation to survive and then mate to produce the offspring of the next generation.

For the most part, the formulations considered in this chapter deal with one autosomal locus with two alleles so that there are only three genotypes under consideration for each sex, which in turn implies that it is sufficient to consider 9 types of couples or sexual partnerships. From the point of view of stochastic processes, the class of stochastic process under consideration in this chapter are self regulating multitype branching processes, which are extensions of those introduced and applied in chapters 9 and 10. When only three genotypes are under consideration, the state space $\mathfrak{S}$ of the process is the set of all 6-dimensional vectors of non-negative integers of the form $\boldsymbol{w} = (x_1, x_2, x_3, y_1, y_2, y_3)$, where the first three components denote the numbers of females for each of the three genotype and the last three components are defined similarly for males. In such Markov chains, the state $\boldsymbol{w} = \mathbf{0}$ is an absorbing state under the assumptions underlying most of the computer simulation experiments reported in this chapter. As is often observed in these experiments, when the results of a Monte Carlo simulation experiment are summarized statistically, the evidence suggests that the process has converged to some kind of stationary distribution, which can in theory be attributed to convergence to the quasi-stationary distribution of the process, given that extinction has not occurred. Although most of this chapter deals with formulations involving one autosomal locus with two alleles, in the last section of the chapter some issues are discussed that will arise when attempts are made to the extend the two sex process under consideration to the case of two linked autosomal loci with two alleles at each locus.

There is an extensive literature on population genetics that deals with models of mutation and selection. A review of this literature, based on many references, may be found in the text book Hartl and Clark (1989) and also later editions of this book. For the most part, the models of mutation and selection described in this book are deterministic in nature and are based on theories of gene or gamete frequencies with selection characterized in term of viabilities or notions of fitness. When models of this type are tractable mathematically, formulas for fixed points of non-linear difference equations are often exhibited, but little attention is paid to rates of convergence to these fixed points in computer experiments.

In contrast to many of models of selection and mutation in classical population genetics, all models described in this chapter are stochastic and attention is focused on the demographics of a population, described in terms of counts of the number of individuals of each genotype, as it evolves from small founder populations rather than on gene or gametic frequencies. Furthermore, deterministic models are embedded in stochastic processes so that the performances of the stochastic model and the embedded deterministic model may be compared in computer simulation experiments. Another advantage of using the class of branching process under consideration is that allows an investigator to study the transient evolutionary behavior of the process before convergence to a quasi-stationary distribution. As will be shown in the experiments reported in this chapter, this transient period of evolution often involves a high level of stochasticity of the process, which is of interest in applications of the model to actual populations. In the book by Durrett (2008) some attention is also given to studying the effects demographic factors in the evolution of populations, but in that book the retrospective perspective is emphasized in the sense that these effects are described in terms of models of coalescence. In this chapter, however, attention will be focused on the forward-in-time perspective, which is much easier to deal with from both the mathematical and computational points of view.

11.2 Gametes, Genotypes and Couple Types in a Two Sex Stochastic Population Process

The purpose of this section is to set down some notation and concepts that will be utilized in this and subsequent sections of this chapter to develop a two sex stochastic population process that may be used to study the impact of mutation and selection on the evolution of a diploid population. In a word, the class of two sex stochastic processes with discrete time generations to be studied in this chapter will be an extension of the multitype self regulating branching processes formulated in chapter 9 and studied in chapter 10 by applying Monte Carlo simulation methods. Let $\mathfrak{G}$ denote a finite set of gametes with respect to the set of autosomal chromosomes. In man, for example, there are 22 pairs of autosomal chromosomes and in this case the set of gametes under consideration would consist of egg or sperm cells containing 22 chromosomes which arise from a type of reduction cell division called meiosis. Moreover, each gamete contains only one member of each homologous pair of chromosomes and the production of gametes resembles a type of random process discussed in chapter 2. Elements of the set $\mathfrak{G}$ will be denoted by symbols of the form $i, j, k \cdots$ with or without subscripts. In applications of ideas being developed in this section, only a small subset of the set $\mathfrak{G}$ of autosomal gametes will be considered in any computer experiment.

A diploid genotype will be denoted by the ordered pair (i, j), where $i \in \mathfrak{G}$ is the gamete contributed by the female parent and $j \in \mathfrak{G}$ is the gamete contributed by the male parent. Let $\mathfrak{T}_f = ((i, j) \mid i \in \mathfrak{G}, j \in \mathfrak{G})$ denote the set of all genotypes in the female population in some generation

$n = 0, 1, 2, \ldots$ and define the set $\mathfrak{T}_m$ similarly for males. Elements of the set $\mathfrak{T}_f$ will be denoted by the symbol τ_f, and, similarly, elements of the set $\mathfrak{T}_m$ will be denoted by the symbol τ_m. A couple, consisting of a female and male, will be said to be of type $\kappa = ((i,j), (k,l))$ if the female has genotype (i,j) and the male has genotype (k,l). To lighten the notation a couple type may also be denoted by the symbol $\kappa = (\tau_f, \tau_m)$. Let $\mathfrak{K}$ denote the set of all couple types in some generation. In a two sex process, offspring arise only from the matings of couples. Stochastic processes describing the production of offspring by couples will be described in a subsequent section. Females and males who are not members of a couple will be called singles.

The next step in the formulation of a two sex population process is that of defining some random functions for every generation $n = 0, 1, 2, \ldots$, taking values in the set $(x \mid x = 0, 1, 2, \ldots)$ of nonnegative integers, that will play a fundamental role in the computer implementation of the process. For every genotype $\tau_f \in \mathfrak{T}_f$, let the random function $X(n; \tau_f)$ denote the number of single females of genotype τ_f in the population in generation n. Similarly, for every genotype $\tau_m \in \mathfrak{T}_m$, let the random function $Y(n; \tau_m)$ denote the number of single males in the population of genotype τ_m in generation n. Lastly, let the random function $Z(n; \kappa)$ denote the number of couples of type $\kappa = (\tau_f, \tau_m) \in \mathfrak{K}$ in the population in generation n.

As couples are formed from pairs of single females and males, the array of couple types must satisfy a set in inequalities with probability one. Because the number of females in couples of genotype τ_f cannot exceed the number of single females of genotype τ_f from which they were formed, it follows that the inequality

$$\sum_{\tau_m \in \mathfrak{T}_m} Z(n; \tau_f, \tau_m) \leq X(n; \tau_f) \tag{11.2.1}$$

must hold with probability one for every $\tau_f \in \mathfrak{T}_f$ and generation n. Similarly, for males in couples of genotype τ_m, the inequality

$$\sum_{\tau_f \in \mathfrak{T}_f} Z(n; \tau_f, \tau_m) \leq Y(n; \tau_m) \tag{11.2.2}$$

must hold with probability one for every genotype $\tau_m \in \mathfrak{T}_m$ and generation n. In the next section, a couple formation process will be presented that has the property that inequalities (11.2.1) and (11.2.2) will hold with probability one.

11.3 The Parameterization of Couple Formation Processes

A fundamental problem in the formulating a stochastic two sex evolutionary model of population dynamics with couple formation is that of finding procedures for simulating realizations of the random variables $Z(n;\kappa)$ for all $\kappa \in \mathfrak{K}$ such that the inequalities in (11.2.1) and (11.2.2) hold with probability one. One approach to this problem was described in Mode (1995) in which a procedure consisting of a mixture of bivariate distributions was used to construct a couple formation process. As it turned out realizations of the collection of random variables $(Z(n;\kappa) \mid \kappa \in \mathfrak{K})$ computed from this procedure did not automatically satisfy these inequalities, but by the use of a rejection method different realizations of the couple formation process were computed until one was found that satisfied inequalities (11.2.1) and (11.2.2). Even though it was possible to write software to implement this procedure with acceptable computer execution times, one was left with a feeling that it would be preferable if a procedure could be found and implemented such that the inequalities in question were automatically satisfied with probability one.

While working on a stochastic model describing the evolution of a HIV/AIDS epidemic in a population of heterosexuals with couple formation, a procedure was found such that the desired inequalities would be satisfied with probability one, see Mode and Sleeman (2000) section 12.4. The purpose of this section is to recast this formulation into a form that will be usable in the context of the population genetic process under consideration. To this end, let the random functions $X(n;\tau_f)$ and $Y(n;\tau_m)$, for $\tau_f \in \mathfrak{T}_f$ and $\tau_m \in \mathfrak{T}_m$, denote the numbers of single females and males present in a population in generation n. Given these numbers of single females and males, let the random function $N_C(n;\kappa)$, for $\kappa \in \mathfrak{K}$, denote the potential number of couples of type $\kappa = (\tau_f, \tau_m)$ that may be formed in generation n from single females and males. A useful way of thinking about this random function is that it represents the maximum number of couples of type κ that could be formed, given the interaction of single females and males in their searches for mates in generation n.

In the formulation under consideration, the actual number $Z(n;\kappa)$ of couples of type κ formed in generation n will be viewed as a realization of $N_C(n;\kappa) \geq 0$ conditionally independent Bernoulli trials with some common probability $p(\kappa)$ that will be discussed subsequently. Under this view, the inequalities $0 \leq Z(n;\kappa) \leq N_C(n;\kappa)$ for all $\kappa \in \mathfrak{K}$ hold with probability one. Therefore, to prove that the random functions $(Z(n;\kappa) \mid \kappa \in \mathfrak{K})$ satisfy the desired inequalities, it will suffice to prove that the random functions $(N_C(n;\kappa) \mid \kappa \in \mathfrak{K})$ satisfy these inequalities with probability one.

A basic component of the couple formation process under consideration is that of social contact probabilities of single females and males in their searches for mates. Given a female of type $\tau_f \in \mathfrak{T}_f$ in generation n, let $\gamma_f(n;\tau_f,\tau_m)$ denote the conditional probability that she has a contact with a male of type $\tau_m \in \mathfrak{T}_m$. Similarly, given a male of type $\tau_m \in \mathfrak{T}_m$ in generation n, let $\gamma_m(n;\tau_m,\tau_f)$ denote the conditional probability that he has contact with a female of type $\tau_f \in \mathfrak{T}_f$. In what follows, these contact probabilities will be constructed as functions of, among other things, the frequencies of genotypes of both sexes in generation n. In this construction, all these probabilities will lie in the closed interval $[0,1]$, and, those for females, for example, will satisfy the condition

$$\sum_{\tau_m \in \mathfrak{T}_m} \gamma_f(n;\tau_f,\tau_m) = 1 \tag{11.3.1}$$

for all $\tau_f \in \mathfrak{T}_f$ and generations n. An analogous condition will hold for the contact probabilities for males. Let

$$\boldsymbol{\gamma}_f(n;\tau_f) = \left(\gamma_f(n;\tau_f,\tau_m) \mid \tau_m \in \mathfrak{T}_m\right) \tag{11.3.2}$$

denote a vector of contact probabilities for single females of type $\tau_f \in \mathfrak{T}_f$, and let $\boldsymbol{\gamma}_m(n;\tau_m,\tau_f)$ denote a similar vector of contact probabilities for single males of type $\tau_m \in \mathfrak{T}_m$ in generation n.

For single females of type τ_f in generation n, let the random function $Z_f(n;\tau_f;\tau_m)$ be the number of single males of type τ_m selected as potential sexual partners, and let

$$\boldsymbol{Z}_f(n;\tau_f) = (Z_f(n;\tau_f;\tau_m) \mid \tau_m \in \mathfrak{T}_m) \tag{11.3.3}$$

denote of vector of these random functions. Given the number $X(n;\tau_f)$ of single females of τ_f in generation n, it will be assumed that the random vector $\boldsymbol{Z}_f(n;\tau_f)$ has a conditional multinomial distribution with index $X(n;\tau_f)$ and probability vector $\boldsymbol{\gamma}_f(n;\tau_f)$. In symbols,

$$\boldsymbol{Z}_f(n;\tau_f) \sim CMultinom\left(X(n;\tau_f), \boldsymbol{\gamma}_f(n;\tau_f)\right). \tag{11.3.4}$$

Similarly, let $\boldsymbol{Z}_m(n;\tau_m)$ denote the corresponding vector of random functions for males of type τ_m in generation n. Then, it will also be assumed that

$$\boldsymbol{Z}_m(n;\tau_m) \sim CMultinom\left(Y(n;\tau_m), \boldsymbol{\gamma}_m(n;\tau_m)\right). \tag{11.3.5}$$

Since the number of pair-wise contacts of type (τ_f, τ_m) cannot exceed the number of single females seeking single males of type τ_m as well as the number of single males of type τ_m seeking single females of type τ_f, it follows that a plausible choice for the random function $N_C(n; \tau_f, \tau_m)$ is

$$N_C(n; \tau_f, \tau_m) = \min\left(Z_f(n; \tau_f, \tau_m), Z_m(n; \tau_m, \tau_f)\right) \qquad (11.3.6)$$

for all $\kappa = (\tau_f, \tau_m) \in \mathfrak{K}$ and generations n. Moreover, from $(11.3.4)$, it follows that

$$\sum_{\tau_m \in \mathfrak{T}_m} Z_f(n; \tau_f, \tau_m) = X(n; \tau_f) \qquad (11.3.7)$$

with probability one for all generations n and types $\tau_f \in \mathfrak{T}_f$. From $(11.3.6)$ and $(11.3.7)$, it follows that

$$\sum_{\tau_m \in \mathfrak{T}_m} N_C(n; \tau_f, \tau_m) \leq \sum_{\tau_m \in \mathfrak{T}_m} Z_f(n; \tau_f, \tau_m) = X(n; \tau_f) \qquad (11.3.8)$$

with probability one for all generations n and types $\tau_f \in \mathfrak{T}_f$. By a similar argument, it can be shown that

$$\sum_{\tau_f \in \mathfrak{T}_f} N_C(n; \tau_f, \tau_m) \leq \sum_{\tau_f \in \mathfrak{T}_f} Z_m(n; \tau_m, \tau_f) = Y(n; \tau_m) \qquad (11.3.9)$$

with probability one for all generations n and types $\tau_m \in \mathfrak{T}_m$. We have thus shown that the collection of random functions $(N_C(n; \kappa) \mid \kappa \in \mathfrak{K})$ satisfy inequalities (11.2.1) and (11.2.2) with probability one for every generation n.

The next step in the formulation of the couple formation process is that of setting up a general procedure for calculating contact probabilities for single females and males. Given a single female of type τ_f in any generation n, let $\alpha_f(\tau_f, \tau_m)$ denote the conditional probability she finds a single male of type τ_m acceptable as a sexual partner. Similarly, given a single male of type τ_m in any generation n, let $\alpha_m(\tau_m, \tau_f)$ denote the conditional probability he finds a single female of type τ_f acceptable as a sexual partner. These acceptance probabilities will be discussed in more detail subsequently.

By definition, the frequency of type $\tau_f \in \mathfrak{T}_f$ in the single population in generation n is

$$U_f(n; \tau_f) = \frac{X(n; \tau_f)}{X(n; \circ)}, \qquad (11.3.10)$$

where

$$X(n; \circ) = \sum_{\tau_f \in \mathfrak{T}_f} X(n; \tau_f) \qquad (11.3.11)$$

and $X(n;\circ) > 0$. If $X(n;\circ) = 0$, then $U_f(n;\tau_f) = 0$. Let $U_m(n;\tau_m)$ denote the relative frequency of single males of type $\tau_m \in \mathfrak{T}_m$ in population in generation n, which is defined in the same way as that for single females.

By the law of total probability, the probability that in generation n a single female of type τ_f has contact with some single male is

$$\sum_{\tau_m \in \mathfrak{T}_m} U_m(n;\tau_m)\, \alpha_f(\tau_f,\tau_m), \tag{11.3.12}$$

and, by an application of Bayes' formula, it follows that the conditional probability that this contact is with a single male of genotype τ_m is

$$\gamma_f(n;\tau_f,\tau_m) = \frac{U_m(n;\tau_m)\, \alpha_f(\tau_f,\tau_m)}{\sum_{\tau_m \in \mathfrak{T}_m} U_m(n;\tau_m)\, \alpha_f(\tau_f,\tau_m)}. \tag{11.3.13}$$

Similarly, the conditional probability that a single male of genotype τ_m has contact with a single female of genotype τ_f is

$$\gamma_m(n;\tau_m,\tau_f) = \frac{U_f(n;\tau_f)\, \alpha_m(\tau_m,\tau_f)}{\sum_{\tau_f \in \mathfrak{T}_f} U_f(n;\tau_f)\, \alpha_m(\tau_m,\tau_f)}. \tag{11.3.14}$$

From these equations, it is easy to see that if there are constants a and b in the interval $(0,1)$ such that $\alpha_f(\tau_f,\tau_m) = a$ and $\alpha_m(\tau_m,\tau_f) = b$ for all contact types (τ_f,τ_m) and (τ_m,τ_f), then

$$\gamma_f(n;\tau_f,\tau_m) = U_m(n;\tau_m) \tag{11.3.15}$$

and

$$\gamma_m(n;\tau_m,\tau_f) = U_f(n;\tau_f). \tag{11.3.16}$$

When these conditions hold, the selection of potential sexual partners by single females and males is, by definition, random.

When the numbers of genotypes in the sets $\mathfrak{T}_f$ and $\mathfrak{T}_m$ are large, the problem of specifying numerical values for the acceptance probabilities become problematic. However, in those cases in which it is feasible to associate a numerical value in the form of a score or measurement with each genotype, then the probabilities in question can be parameterized with relative ease. For any $\tau_f \in \mathfrak{T}_f$, let x_{τ_f} denote the numerical value assigned to genotype τ_f. Similarly, let y_{τ_m} denote the numerical value assigned to the male genotype $\tau_m \in \mathfrak{T}_m$. Then, one approach to parameterizing the acceptance probabilities is to suppose they have the forms

$$\alpha_f(\tau_f,\tau_m) = \exp\left(-\beta_f \mid x_{\tau_f} - y_{\tau_m} \mid\right) \tag{11.3.17}$$

and

$$\alpha_m(\tau_m,\tau_f) = \exp\left(-\beta_m \mid y_{\tau_m} - x_{\tau_f} \mid\right), \tag{11.3.18}$$

where the betas are parameters such that $\beta_f \geq 0$ and $\beta_m \geq 0$. Observe that with this parameterization, if $\beta_f = \beta_m = 0$, then the selection of potential sexual partners would be random. Moreover, from these formulas, it also follows that the greater the distance between two potential sexual partners, the smaller is the acceptance probability. It is also of interest to note that the probability $p(\kappa)$ for couple of type $\kappa = (\tau_f, \tau_m)$ in the formulation for calculating the actual number of couples $Z(n; \kappa)$ formed from $N_C(n; \kappa)$ potential couples of type κ in generation n may be parameterized in a similar manner. For in this case one could choose a parameter β_C such that $\beta_C \geq 0$ and let

$$p(\kappa) = \exp\left(-\beta_C \mid x_{\tau_f} - y_{\tau_m} \mid\right) \tag{11.3.19}$$

for all $\kappa \in \mathfrak{K}$.

It is recognized that the process describing couple formation just described is not sufficiently general to cover all mating systems that may exist in nature. For example, the process under consideration does not take into account polygamy in which females and males may have more than one sexual partner. But, on the other hand, it seems to capture, to some extent, part of a process in which the self interests of females and males, as expressed in terms of acceptance probabilities, lead to monogamous partnerships that result in the production of offspring.

11.4 An Example of Couple Formation Process with Respect to an Autosomal Locus with Two Alleles

In some cases, when there are relatively few genotypes under consideration, the acceptance probabilities may be defined with relative ease. Such cases arise when models with respect to one autosomal locus with two alleles are under consideration and there is no mutation. Let A and a denote two alleles at some autosomal locus. Then, the set of genotypes for the female population is

$$\mathfrak{T}_f = (AA, Aa, aa), \tag{11.4.1}$$

and the set $\mathfrak{T}_m$ of genotypes for the male population is the same as that for females. To lighten the notation, these three genotypes will be numbered 1,2 and 3 and suppose allele A is dominant to a. Then, among the three genotypes there are only two phenotypes, because the genotypes AA and Aa express the dominant phenotype and the genotype aa expresses the recessive phenotype.

Now suppose that potential sexual partners are chosen by phenotype, and let

$$\boldsymbol{A}_f = \begin{pmatrix} \alpha_f(1,1) & \alpha_f(1,2) & \alpha_f(1,3) \\ \alpha_f(2,1) & \alpha_f(2,2) & \alpha_f(2,3) \\ \alpha_f(3,1) & \alpha_f(3,2) & \alpha_f(3,3) \end{pmatrix} \tag{11.4.2}$$

denote the 3×3 matrix of acceptance probabilities for single females. By way of an explanation, the rows of the matrix represent that genotypic classification of the single females and the columns the genotypic classifications of the single males.

In general there are nine parameters in this matrix, but when potential sexual partners are chosen according to phenotypes of the females and males, it suffices to consider only three acceptance probabilities for females. Let α_{f1} denote the acceptance probability for females with the dominant phenotype when the males are also express the dominant phenotype. Let α_{f2} denote a acceptance probability of females of the dominant phenotype when the male expresses the recessive phenotype. For the sake of simplicity, it will also be assumed that the probability α_{f2} is in force when the female expresses the recessive phenotype and the male expresses the dominant phenotype. Finally, if both the female and male express the recessive phenotype, let α_{f3} denote the acceptance probability for this case. Then, the 3×3 matrix acceptance probabilities for single females takes the three-parameter form

$$\boldsymbol{A}_f = \begin{pmatrix} \alpha_{f1} & \alpha_{f1} & \alpha_{f2} \\ \alpha_{f1} & \alpha_{f_1} & \alpha_{f2} \\ \alpha_{f2} & \alpha_{f2} & \alpha_{f3} \end{pmatrix}. \tag{11.4.3}$$

Similarly, let α_{m1}, α_{m2} and α_{m3} denote analogous acceptance probabilities for single males. Then the 3×3 matrix $\boldsymbol{A}_m$ of acceptance probabilities for single males would have the same form as that for $\boldsymbol{A}_f$. Altogether the couple formation process under consideration depends on six parameters, which could be accommodated in a software implementation of the process with relative ease. Indeed, it seems advisable to provide for six parameters in the software, for, if an investigator wished to consider the case $\boldsymbol{A}_f = \boldsymbol{A}_m$, he or she would be free to do so, because a six parameter model would be accommodated in the software.

It is tempting to interpret matrices of the displayed in (11.4.3) as payoff matrices of evolutionary game theory, see Nowak (2006) for discussion of evolutionary games and the investigators who have worked in this field. Briefly, in these discussions, the idea of fitness of a type is equated with means of the elements of a payoff matrix. As an illustration of this idea, suppose there are two types 1 and 2 and the pay off matrix has the form

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{21} \\ a_{21} & a_{22} \end{pmatrix}. \tag{11.4.4}$$

Let x_1 and x_2 denote the frequencies of the two types in the population, where $x_1 \in [0,1]$ and $x_1 + x_2 = 1$. Let $\boldsymbol{x} = (x_1, x_2)$ denote a vector of these frequencies. Then, the fitness of a type 1 individual in the formulation presented by Nowak is defined as the weighted mean

$$f_1(\boldsymbol{x}) = x_1 a_{12} + x_2 a_{21} \tag{11.4.5}$$

and the fitness of an individual of type 2 defined similarly. In the two sex evolutionary process under consideration, however, matrices of form (11.4.3) will be viewed merely as one of the components of natural selection along with two other components; namely the expected number of offspring produced by each couple type and the probabilities of their female and male offspring survive to reproduce the next generation.

11.5 Genetics and Offspring Distributions

In the class of two sex processes under consideration, describing the evolution of a diploid species such as man, genes are passed from parents to offspring by those couples who reproduce. For example, consider a couple of type $\kappa = (\tau_f, \tau_m) \in \mathfrak{K}$ in which the female is of genotype $\tau_f = (i,j)$ and the male is of genotype $\tau_m = (k,l)$, where the gamete on the left is that received from the female parent and that on the right is the one received from the male parent. The principal objective of this section is to describe the derivation of the offspring distribution for each couple type $\kappa = (\tau_f, \tau_m)$, which consists of probabilities of the form $p(\kappa;\tau)$, where $p(\kappa;\tau)$ is the conditional probability that a couple of type κ produces an offspring of genotype τ. Initially, the sex of this offspring will be ignored but it will be incorporated into the notation as the development of ideas progresses.

To illustrate the basic concepts, the simple case of one autosomal locus will be considered with respect to two alleles denoted by A and a as discussed in section 11.4, and it will be supposed that allele A may mutate to allele a with probability μ_{12} and allele a may mutate to allele A with probability μ_{21}. When this assumptions are in force, the 2×2 mutation matrix takes the form

$$\mathfrak{M} = \begin{pmatrix} \mu_{11} & \mu_{12} \\ \mu_{21} & \mu_{22} \end{pmatrix}, \tag{11.5.1}$$

where the diagonal elements are chosen such that each row of the matrix sums for 1.

The first step in the procedure for calculating the offspring distribution for each couple type is to derive formulas for the production of gametes under the assumption that mutations may occur for each of the three genotypes being considered. Presented below is a table in which the gamete distributions are expressed symbolically in terms of mutation probabilities for each of the three genotypes under consideration. To interpret this table, consider the gametes produced by genotype AA is the second row of this table. The probability that the left allele in genotype AA is transmitted to gametic pool of this genotype and a gamete is of type A is $\mu_{11}/2$. Similarly for the right allele in the genotype AA this probability is also $\mu_{11}/2$. Therefore, the probability the genotype AA produces a gamete of type A is $\mu_{11}/2 + \mu_{11}/2 = \mu_{11}$. In general, a similar rationale was used to derive the other formulas in the table. It should be noted the rows 2,3, and 4 of this table, when summed over columns 2 and 3, are all 1.

Table 11.5.1 Gamete Distributions for the Three Genotypes

$\cdot$	***A***	***a***
AA	μ_{11}	μ_{12}
Aa	$\frac{1}{2}(\mu_{11}+\mu_{21})$	$\frac{1}{2}(\mu_{12}+\mu_{22})$
aa	μ_{21}	μ_{22}

From table 11.5.1, observe that the probabilities that a mating of type $AA \otimes AA$ will produce offspring with genotypes AA, Aa and aa are $\mu_{11}^2, 2\mu_{11}\mu_{12}$ and μ_{12}^2, respectively. By continuing in this manner, it would be possible to derive formulas for each offspring distribution associated with each of the nine types of matings but the details will be left to an interested reader. It is, however, of basic interest to describe a general method for calculating the offspring distribution by genotype for any mating of type $\kappa = (\tau_f, \tau_m)$. Let $\mathfrak{G}_g$ denote the set of possible gametes with respect to an autosomal locus under consideration, and $p_g(\tau_f; \nu)$ denote the probability that a female of genotype τ_f produces a gamete of type $\nu \in \mathfrak{G}_g$. Similarly, let $p_g(\tau_m; \nu')$ denote the probability that a male of genotype τ_m produces a gamete of type $\nu' \in \mathfrak{G}_g$. Then, the probability that a couple of type $\kappa = (\tau_f, \tau_m)$ produces an offspring of genotype $\tau = (\nu, \nu')$ may be computed as

$$p\left(\kappa;\tau\right)=p_g\left(\tau_f;\nu\right)p_g\left(\tau_m;\nu'\right) \tag{11.5.2}$$

for all couple types κ and genotypes $\tau = (\nu, \nu')$. Given a table or array as shown in Table 11.5.1 containing numerical versions of gamete distribution for each genotype, software may be written to calculate the offspring distribution for each couple type by implementing formula (11.5.2).

To accommodate the sex of each offspring in the formulation under consideration, let p_f denote the probability that an offspring is female and let $p_m = 1 - p_f$ denote the probability an offspring is male, where $0 < p_f < 1$. Next consider the vector of probabilities

$$\boldsymbol{p}\left(\kappa\right)=\left(p\left(\kappa;\tau\right)\mid\tau\in\mathfrak{T}\right), \tag{11.5.3}$$

where $\mathfrak{T}$ is the set of genotypes under consideration for both sexes. For the case of an autosomal locus under consideration this set consists of the three genotypes, AA, Aa, aa. The sex of an offspring will be indicated by the numbers 1 and 2, where 1 indicates an offspring is female and 2 indicates an offspring is male. Given these indicators, let $q\left(\kappa;1,\tau\right)=p_f\ p\left(\kappa;\tau\right)$ denote the conditional probability that an offspring of a couple of the type $\kappa\in\mathfrak{K}$ is female and of genotype τ. The probability $q\left(\kappa;2,\tau\right)$ is defined similarly for male offspring of a couple of type κ.

To describe the production of female and male offspring for each couple type $\kappa\in\mathfrak{K}$ with given genotypes, it will be helpful to define two vectors. Let

$$\mathbf{q}\left(\kappa;1\right)=\left(q\left(\kappa;1,\tau\right)\mid\tau\in\mathfrak{T}\right) \tag{11.5.4}$$

denote a vector of probabilities of genotypes for female offspring and define the vector $\mathbf{q}\left(\kappa;2\right)$ similarly for male offspring. Then, the offspring distribution for a couple of type κ may be represented in the form of the partitioned vector

$$\boldsymbol{q}\left(\kappa\right)=\left(\mathbf{q}\left(\kappa;1\right),\mathbf{q}\left(\kappa;2\right)\right), \tag{11.5.5}$$

where the components of the vector denote the probabilities of female and male offspring by sex and genotype. For the case of the autosomal locus under consideration with two alleles, this vector would have six components. In what follows in subsequent sections, the symbols p_f and p_m suppressed to simplify the notation and the elements of the vector $\boldsymbol{q}\left(\kappa\right)$ will be denoted by symbols of the form $q\left(\kappa;k,\tau\right)$, where $k = 1, 2$ and $\tau\in\mathfrak{T}$.

To accommodate a component of natural selection, let the random variable $N\left(\kappa\right)$, which takes values in the set of non-negative integers, denote the total number of offspring produced by a couple of type $\kappa \in \mathfrak{K}$ per generation. As in the foregoing chapters, it will be assumed that this random variable has a Poisson distribution with parameter $\lambda(\kappa) > 0$ for every couple type $\kappa \in \mathfrak{K}$. For each mating type $\kappa \in \mathfrak{K}$, let the random variable $W(\kappa; \nu, \tau)$ denote the number of offspring of genotype τ and sex $\nu = 1, 2$ produced by mating type κ per generation, and let

$$\boldsymbol{W}(\kappa) = (W(\kappa; \nu, \tau) \mid \nu = 1, 2; \tau \in \mathfrak{T}) \tag{11.5.6}$$

be a vector of these random variables. Then, given a value of $N(\kappa)$, it will be assumed that the vector $\boldsymbol{W}(\kappa)$ has a conditional multinomial distribution with index $N(\kappa)$ and probability vector $\boldsymbol{q}(\kappa)$. In symbols,

$$\boldsymbol{W}(\kappa) \sim CMultinom(N(\kappa), \boldsymbol{q}(\kappa)). \tag{11.5.7}$$

For the case of one autosomal locus with two alleles under consideration, the set $\mathfrak{K}$ of couple types contains 9 elements. For example, if the three genotypes AA, Aa,aa are numbered 1,2,3, then the Poisson parameters for the offspring distributions may be represented in the form of a 3×3 matrix

$$\begin{pmatrix} \lambda(1,1) & \lambda(1,2) & \lambda(1,3) \\ \lambda(2,1) & \lambda(2,2) & \lambda(2,3) \\ \lambda(3,1) & \lambda(3,2) & \lambda(3,3) \end{pmatrix}. \tag{11.5.8}$$

By way of interpreting this matrix, the symbol $\lambda(1,1) > 0$ is the expected number of offspring produced by a mating of type $\kappa = (1,1)$ with the explicit form $AA \otimes AA$, indicating that a female of genotype AA is mated with a male of genotype AA and the other entries in the matrix have similar interpretations. It is clear that the model in its most general form has 9 parameters, but by making some plausible simplifying assumptions, this number can be reduced. If, for example, it is assumed that this matrix is symmetric so that $\lambda(i,j) = \lambda(j,i)$ for all pairs such that $i \neq j$, then there would be only six parameters to specify. Observe the condition $\lambda(i,j) = \lambda(j,i)$ implies that the expected number of offspring from a mating in which the female is of genotype i and the male is of genotype j is the same for females of genotype j mated with males of genotype i.

This number of parameters can be further reduced for the case allele A is dominant to allele a and the reproductive success of mating depends on the phenotypes of the parents. In such situations, all matings of the form $A - \otimes A-$, indicating that both the female and male express phenotype A, could be assigned the parameter values λ_1. Similarly, matings of the type $aa \otimes aa$ would be assigned the parameter value λ_2. But for mating types in which the parents have different phenotypes, a third parameter value λ_3

would be assigned. Under these assumptions the matrix in (11.5.8) would have the form

$$\begin{pmatrix} \lambda_1 & \lambda_1 & \lambda_3 \\ \lambda_1 & \lambda_1 & \lambda_3 \\ \lambda_3 & \lambda_3 & \lambda_2 \end{pmatrix}. \tag{11.5.9}$$

By using various ordering of these three parameters, one could characterize the expected reproductive success of each of the nine types of matings. For example, if these parameters had the ordering $\lambda_3 > \lambda_1 > \lambda_2 > 0$, then those matings in which the parents had different phenotypes would on average have the greatest reproductive success and the reproductive success of matings of the phenotype $A - \otimes A-$ would be greater than those of type $aa \otimes aa$.

11.6 Overview of a Self-Regulating Population Process

The purpose of this section is to describe the Monte Carlo simulation algorithms that will be utilized in simulating realizations of the two sex population process under consideration. For the sake of concreteness, it will be supposed that some autosomal locus is under consideration. In generation $n \geq 0$, let $X(n;\tau_f)$ and $Y(n;\tau_m)$ denoted the number of single females and males of genotypes τ_f and τ_m, respectively, who have survived from the preceding generation and begin the search for mates. Given these numbers of single females and males in generation n, let the random function $Z(n;\kappa)$, where $\kappa = (\tau_f, \tau_m)$, denote the total number of couples of type κ formed in generation n. Recall that the procedures for calculating realizations of the random functions in the class $(Z(n;\kappa) \mid \kappa \in \mathfrak{K})$ were described in section 11.3.

In generation n, let the random function $T(n;\kappa;k,\tau)$ denote the total number of offspring of type (k,τ) produced by couples of type κ, where $k = 1, 2$ denotes the sex and $\tau \in \mathfrak{T}$ the genotype of the offspring. For $Z(n;\kappa) > 0$ and $\nu = 1, 2, \ldots, Z(n;\kappa)$, let $W_\nu(\kappa;k,\tau)$ denote a collection of conditionally independent random variables, given a values of $Z(n;\kappa)$. For a definition of this random variable see (11.5.6). Then the random function $T(n;\kappa;k,\tau)$ is given by the random sum

$$T(n;\kappa;k,\tau) = \sum_{\nu=1}^{Z(n;\kappa)} W_\nu(\kappa;k,\tau), \tag{11.6.1}$$

where $T(n;\kappa;k,\tau)=0$ if $Z(n;\kappa)=0$. Let the random function $V(n;k,\tau)$ denote the total number of all offspring of type (k,τ) produced in generation n. Then, this random function is given by the sum

$$V(n;k,\tau)=\sum_{\kappa\in\mathfrak{K}}T(n;\kappa;k,\tau). \tag{11.6.2}$$

In particular, $V(n;1,\tau)$ is the number of females of genotype τ and $V(n;2,\tau)$ is the number of males among the offspring of genotype τ. Therefore, the total number of females among the offspring of couples in generation n is

$$V(n;1,\circ)=\sum_{\tau\in\mathfrak{T}}V(n;1,\tau), \tag{11.6.3}$$

and the sum $V(n;2,\circ)$ for males is defined similarly. From these definitions, it follows that the total number of offspring produced by couples in generation n is given by

$$T(n)=V(n;1,\circ)+V(n;2,\circ). \tag{11.6.4}$$

To take into account that all offspring produced in generation n may not survive to form couples and produce offspring in generation $n+1$, it will be assumed that the probability that any individual of type (k,τ) survives depends on the random variable $T(n)$. Let $s(n;k,\tau)$ denote the conditional probability that an offspring of generation n survives to participate in the reproduction, given $T(n)$. Then, it will be assumed that this function has the parametric of a Weibull survival function

$$s(n;k,\tau)=\exp\left(-(\beta(k,\tau)T(n))^{\alpha(k,\tau)}\right), \tag{11.6.5}$$

where $\alpha(k,\tau)$ and $\beta(k,\tau)$ are positive parameters defined for all types (k,τ) of offspring.

Having defined and parameterized survival probabilities for each type of offspring (k,τ), we are now in a position to set down algorithms for computing realizations of the random functions $X(n+1;\tau_f)$ and $Y(n+1;\tau_m)$, the numbers of female and male offspring that survive to form couples and produce the offspring of generation $n+1$. Given a value of the random function $V(n;1,\circ)$, a realization of the random function $X(n+1;\tau_f)$ is computed as a sample from a conditional binomial distribution with index $V(n;1,\tau_f)$ and probability $s(n;1,\tau_f)$. In symbols,

$$X(n+1;\tau_f)\sim CBinom(V(n;1,\tau_f),s(n;1,\tau_f)) \tag{11.6.6}$$

for each genotype τ_f. Similarly, for each male genotype τ_m, we have

$$Y(n+1;\tau_m)\sim CBinom(V(n;2,\tau_m),s(n;2,\tau_m)). \tag{11.6.7}$$

Given realizations of these random functions for single females and males by genotype, the procedures described in section 11.3 may be applied to compute realizations of random functions in the collection $(Z(n+1;\kappa) \mid \kappa \in \mathfrak{K})$ for generation $n+1$.

In closing, observe that the formulation described in this and previous sections of this chapter accommodates three components of natural selection; namely the choice of mates as characterized by acceptance probabilities for choices of sexual partners, differential reproduction among mating types as characterized by the λ-parameters and differential probabilities of survival by sex and genotype. In subsequent sections of this chapter, the results of computer simulation experiments, designed to study the effects of these three components, will be reported. Actually, because the effects of better competitive ability for some types when compared with other types led to the predominance of the more competitive type as the population evolved in simulation experiments reported in chapter 9, attention in this chapter will be confined to the two other components of selection under consideration in order to avoid the duplication of results that already have been demonstrated in similar experiments.

11.7 Embedding Non-Linear Difference Equations in the Stochastic Population Process

In chapter 9, techniques for embedding deterministic models in a stochastic process were illustrated for both one type and multitype models, and, as was shown in chapters 9 and 10, projecting the evolution of a population according to embedded deterministic model can be informative when compared with a statistically summarized sample of Monte Carlo projections based on algorithms for computing realizations of the stochastic population process. In this section, procedures for embedding non-linear difference equations in a the stochastic process will be described for the two sex population process under consideration. As was illustrated in chapter 9, the idea of embedded deterministic models in a stochastic process is based on a technique of estimating conditional expectations in generation $n+1$, given realizations of the sample functions of the process in generation n. Let the symbol $\Xi(n)$ denote the phrase, given the sample functions of the process in generation n.

Then, the conditional expectations of the random functions $X(n+1;\tau)$ and $Y(n+1;\tau)$, given $\Xi(n)$, are

$$\begin{aligned} E\left[X\left(n+1;\tau\right) \mid \Xi\left(n\right)\right] &= s\left(n;1,\tau\right)V\left(n;1,\tau\right) \\ E\left[Y\left(n+1;\tau\right) \mid \Xi\left(n\right)\right] &= s\left(n;2,\tau\right)V\left(n;2,\tau\right) \end{aligned} \tag{11.7.1}$$

for all $\tau \in \mathfrak{T}$, the set of genotypes for the autosomal locus under consideration. See (11.6.6) and (11.6.7) for a justification of the formulas on the right. The random functions on the right are all non-linear functions of the collection of random functions $\Xi\left(n\right) = \left(X\left(n;\tau\right), Y\left(n;\tau\right) \mid \tau \in \mathfrak{T}\right)$ in generation n, and, therefore, it is not possible to show that the unconditional expectations in (11.7.1) satisfy a set of linear equations. It is, however, possible to obtain estimates of these random functions, given estimates of these random function in generation n. Let the collection $\widehat{\Xi}\left(n\right) = \left(\widehat{X}\left(n;\tau\right), \widehat{Y}\left(n;\tau\right) \mid \tau \in \mathfrak{T}\right)$ denote estimates of these random functions for singles in generation n. Then, from $(11.7.1)$, estimates of the random functions for singles in generation $n+1$ may be computed using the recursive formulas

$$\begin{aligned} \widehat{X}\left(n+1;\tau\right) &= \widehat{s}\left(n;1,\tau\right)\widehat{V}\left(n;1,\tau\right) \\ \widehat{Y}\left(n+1;\tau\right) &= \widehat{s}\left(n;2,\tau\right)\widehat{V}\left(n;2,\tau\right) \end{aligned} \tag{11.7.2}$$

for all genotypes $\tau \in \mathfrak{T}$.

The survival functions $\widehat{s}\left(n;1,\tau\right)$ and $\widehat{s}\left(n;2,\tau\right)$ on the right in (11.7.2) may be estimated by using the collection $\widehat{\Xi}\left(n\right)$ and the formula (11.6.5). Computing values of the estimates $\widehat{V}\ \left(n;k,\tau\right)$ for $k = 1,2$ and $\tau \in \mathfrak{T}$ will, however, be more complicated. First observe that form $(11.5.7)$, it follows that

$$E\left[W\left(\kappa;k,\tau\right)\right] = \lambda\left(\kappa\right)q\left(\kappa;k,\tau\right) \tag{11.7.3}$$

for all $\kappa \in \mathfrak{K}, k = 1,2$ and $\tau \in \mathfrak{T}$. Therefore,

$$E\left[V\left(\kappa;k,\tau\right) \mid \Xi\left(n\right)\right] = \sum_{\kappa\in\mathfrak{K}} Z\left(n;\kappa\right)\lambda\left(\kappa\right)q\left(\kappa;k,\tau\right), \tag{11.7.4}$$

and from this equation, it can be seen that an estimate of the random function $Z\left(n;\kappa\right)$ must be computed, given the collection $\widehat{\Xi}\left(n\right)$. The conditional expectation of $Z\left(n;\kappa\right)$, given $\Xi\left(n\right)$ is

$$E\left[Z\left(n;\kappa\right) \mid \Xi\left(n\right)\right] = N_C\left(n;\kappa\right)p\left(\kappa\right), \tag{11.7.5}$$

see section 11.3 for definitions of the symbols on the right. From this equation, it can be seen that it is necessary to compute an estimate of the random function $N_C\left(n;\kappa\right)$, the potential number of couples of type κ in generation n. See, for example, (11.3.6) for the definition of $N_C\left(n;\kappa\right)$.

To estimate this function observe that from (11.3.4) it follows that

$$E\left[Z_f\left(n;\tau_f,\tau_m\right)\mid\Xi\left(n\right)\right]=X\left(n;\tau_f\right)\gamma_f\left(n;\tau_f;\tau_m\right), \tag{11.7.6}$$

where the contact probability $\gamma_f\left(n;\tau_f;\tau_m\right)$ is computed according to the formula (11.3.13). Hence, an estimate of the random function $Z_f\left(n;\tau_f,\tau_m\right)$ may be computed using the formula

$$\widehat{Z}_f\left(n;\tau_f,\tau_m\right)=\widehat{X}\left(n;\tau_f\right)\widehat{\gamma}_f\left(n;\tau_f;\tau_m\right), \tag{11.7.7}$$

where the symbols on the right denote the functions that have be estimated from the collection of estimates $\widehat{\Xi}\left(n\right)$ for generation n. Similarly, from $(11.3.5)$, it follows that the function $Z_m\left(n;\tau_m,\tau_f\right)$ may be estimated using the formula

$$\widehat{Z}_m\left(n;\tau_m,\tau_f\right)=\widehat{Y}\left(n;\tau_m\right)\widehat{\gamma}_m\left(n;\tau_m,\tau_f\right). \tag{11.7.8}$$

Finally, an estimate of the random function $N_C\left(n;\kappa\right)$ may be computed using the formula

$$\widehat{N}_C\left(n;\kappa\right)=\min\left(\widehat{Z}_f\left(n;\tau_f,\tau_m\right),\widehat{Z}_m\left(n;\tau_m,\tau_f\right)\right) \tag{11.7.9}$$

for every couple type $\kappa=\left(\tau_f,\tau_m\right)$, see formula (11.3.6).

From these results, an estimate of the random function $V(n;k,\tau)$ may be computed using the formula

$$\widehat{V}(n;k,\tau)=\sum_{\kappa\in\mathfrak{K}}\widehat{N}_C\left(n;\kappa\right)p\left(\kappa\right)\lambda\left(\kappa\right)q\left(\kappa;k,\tau\right) \tag{11.7.10}$$

for $k=1,2$ and every genotype $\tau\in\mathfrak{T}$. We thus arrive at the system of non-linear difference equations

$$\begin{aligned}\widehat{X}\left(n+1;\tau\right)&=\widehat{s}\left(n;1,\tau\right)\sum_{\kappa\in\mathfrak{K}}\widehat{N}_C\left(n;\kappa\right)p\left(\kappa\right)\lambda\left(\kappa\right)q\left(\kappa;1,\tau\right)\\ \widehat{Y}\left(n+1;\tau\right)&=\widehat{s}\left(n;2,\tau\right)\sum_{\kappa\in\mathfrak{K}}\widehat{N}_C\left(n;\kappa\right)p\left(\kappa\right)\lambda\left(\kappa\right)q\left(\kappa;2,\tau\right),\end{aligned} \tag{11.7.11}$$

which may be used recursively to estimate the collection of random functions $\left(X\left(n;\tau\right),Y\left(n;\tau\right)\mid\tau\in\mathfrak{T}\right)$ for $n=1,2,\ldots$, given an initial collection of values $\widehat{\Xi}\left(0\right)$ in generation $n=0$. In subsequent sections of this chapter, given assignments of values for all the parameters and the collection $\widehat{\Xi}\left(0\right)$ of assigned initial values for the random functions for singles, equations (11.7.11) will be iterated for generations $n=1,2,3,\ldots,n_1$, where n_1 is some specified positive integer. Moreover, for some fixed number $r\geq 1$ of Monte Carlo replications, realizations of the collection $\Xi\left(n\right)$ of random functions will also be computed and statistically summarized for generations $n=1,2,\ldots,n_1$, using methods similar to those outlined in chapters 9 and 10.

11.8 On the Emergence of a Beneficial Mutation From a Small Founder Population

In experiment 11.8.1, the two sex branching process described in the foregoing sections of this chapter was used to study the emergence of a beneficial mutation when the initial founder population was small. The three genotypes with respect to some autosomal locus with two alleles will be denoted by AA, Aa and aa and it will be assumed that allele A is dominant to allele a. The initial vector for females in generation 0 was chosen as $\boldsymbol{X}_0 = (100, 0, 0)$, indicating the initial population consisted of 100 females of genotype AA. Similarly, the initial population vector for males was chosen as $\boldsymbol{Y}_0 = (105, 0, 0)$. The 2×2 mutation matrix containing the probabilities of mutation among the alleles A and a per generation was chosen as

$$\mathfrak{M} = \begin{pmatrix} \mu_{11} & 10^{-5} \\ 10^{-6} & \mu_{22} \end{pmatrix}, \tag{11.8.1}$$

where the diagonal elements are chosen such that each row of the matrix sums to one. In each Monte Carlo replication of the experiment, the number of generations of evolution was chosen as $G = 2,000$ and the number of Monte Carlo replications of the experiment was chosen as $R = 100$. The time taken to complete this experiment on 3 giga Hertz computer was a little over 8 hours.

All the elements of the 3×3 matrices of acceptance probabilities for females and males, $\boldsymbol{A}_f$ and $\boldsymbol{A}_m$ (see (11.4.1)) were chosen as 1 so that the mating system was, by assumption, random. Similarly, the probabilities of couple formation $p(\kappa)$ for each of the nine couple types $\kappa \in \mathfrak{K}$ were chosen as the constant $p(\kappa) = 0.9$, indicating there was no selection with respect to the formation of couples, see section 11.3 for more details. It was, however, assumed that there was selection with respect to success in reproduction among the couple types. In particular, the matrix of λ-parameters for the Poisson distributions governing the production of offspring was chosen as

$$\boldsymbol{\Lambda} = \begin{pmatrix} 3 & 3 & 3 \\ 3 & 3 & 3 \\ 4 & 4 & 4 \end{pmatrix}. \tag{11.8.2}$$

Briefly, the rationale underlying this choice of reproduction parameters was that it was desired that the mutant genotype aa would eventually appear in the population with high probability, which lead to the assignment of the number 3 for the first two rows of the matrix. It was also assumed the mutant genotype was beneficial in the sense that females with the genotype aa were more successful mothers with respect to the number of offspring they produced on average, which lead to assigning the number 4 in the last row of the matrix. Furthermore, it was assumed that no genotype had a competitive advantage over the others, which led to assigning constant values for the α and β parameters in the Weibull survival functions. In particular, all α-parameters were assinged the value 2 and all β-parameters were assigned the values 10^{-7}. Finally, the probability a child is female was assigned the value $p_f = 100/205$ and the probability a child is a male was assigned the value $p_m = 1 - p_f$.

Figure 11.8.1 contains the graphs of the evolutionary trajectories for the three genotypes in the female population, which were computed by using the embedded deterministic model. As can be seen from this figure, starting from an initial number of 100 females of genotype AA, genotype 1, the trajectory for this genotype climbs to a little less than 2 million individuals in less than 100 generations.

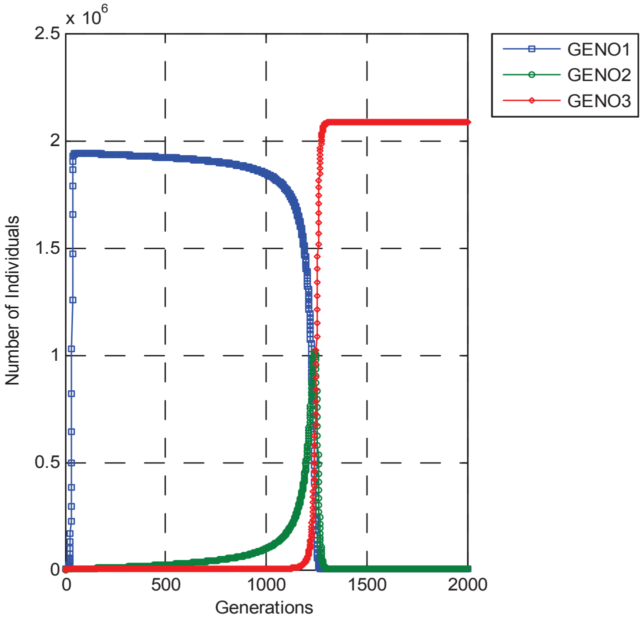

Figure 11.8.1 Graphs for the Trajectories of the Three Genotypes Computed From the Deterministic Model for Experiment 11.8.1.

After reaching a peak at less than 100 generations, the number of individuals of genotype 1 begins a slow decline that lasts for about 1000 generations before it declines precipitously during time span from about 1100 to 1300 generations. This decline in the number of individuals of genotype AA was accompanied by a rapid increase in the number of individuals of the mutant genotype aa, genotype 3, from a small number at about 1200 generations into the projection to over 2 million individuals by about 1300 generations and remains there for the rest of the projection.

In experiment 11.8.1, the rise to predominance of genotype 3 was due to the assumption that selection favored females of genotype aa, because they were more successful in bearing and rearing offspring as indicated in the third row of the matrix in (11.8.2). Interestingly, the number of individuals of the heterozygous genotype Aa, genotype 2, did not appear in the population in detectable numbers until about 600 to 800 generations into the projection and reached a peak at about 1300 generations into the projection and then declined as the genotype aa rose to predominance in the population. The graphs of the trajectories for the three genotypes in the male population were very similar to those for the females population and will, therefore, be omitted.

Presented in Figure 11.8.2 are the graphs of the $MAX, Q50$ and MIN trajectories for females of genotype aa for 2000 generations of evolution, which were computed from 100 Monte Carlo realizations of the stochastic population process under consideration. From this figure it can be seen that the MAX trajectory for genotype aa did not begin to rise sharply until somewhere between 1200 and 1300 generations into the projection of 2000 generations. In this connection, it is interesting to note the MAX trajectory in this figure is very close to the deterministic trajectory for genotype aa in figure 11.8.1.

From an inspection of the $Q50$ and MAX, it may be concluded that there was a considerable amount of variation among the Monte Carlo realizations of the evolutionary process under consideration. For example, the number of generations between the steep rises of the MAX and MIN were separated by nearly 400 generations of evolution, which can be seen by comparing these trajectories in figure 11.8.2.

Moreover, the steep rise in the $Q50$ trajectory for genotype aa did not begin until sometime after generation 1400. This wide spread in the times to steep rises in the three trajectories under consideration is indicative of a high level of stochasticity among realizations of the stochastic population process. However, after the three trajectories cease to rise, the graphs of the

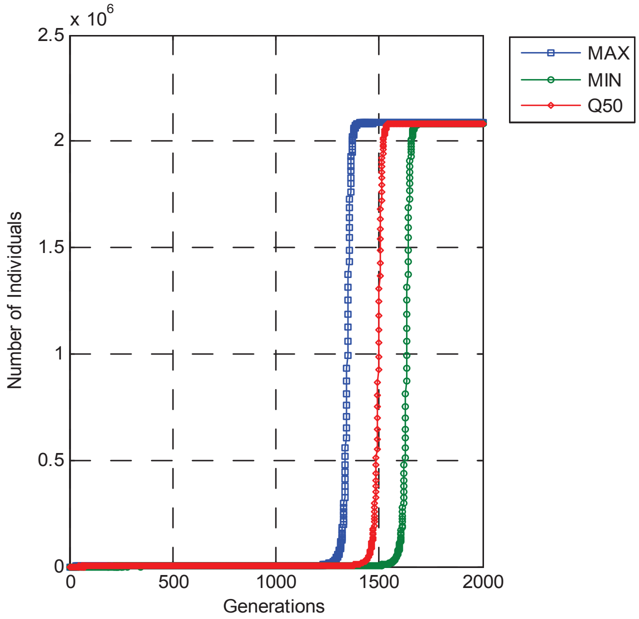

Figure 11.8.2 Graphs of the MAX, MIN and $Q50$ Trajectories for Genotype aa in Experiment 11.8.1.

three trajectories are very close, which is indicative of the process converging to a quasi-stationary distribution. In an effort to conserve space, graphs of the trajectories of the other two genotypes in the female population will be omitted. In passing, it should be mentioned that evolution of the male population with respect to the three genotypes under consideration was very similar to that for females and, thus, will not be shown.

11.9 An Alternative Evolutionary Genetic Model of Inherited Autism

In section 5.5 an evolutionary genetic model for inherited autism was introduced and formulated within the framework of a multitype gamete sample process with constant population size from generation to generation. Briefly, in this formulation attention was focused on gametic frequencies and no attention was given to the numbers of individuals of each of the three autosomal genotypes under consideration. In this section, an alternative evolutionary genetic model of inherited autism will be formulated within the framework of the class of multitype self regulating two sex branching processes with couple formation under consideration in this chapter. As will become apparent in what follows, the framework outlined in the foregoing sections of this chapter can be extended in a straight forward manner to accommodate a theory of inherited autism.

In order to expedite the reading of this section without frequent referrals to section 5.5, a brief outline of a genetic model for inherited autism will be given. It will be supposed that inherited autism is due to mutations that may occur on any of the 22 pairs of autosomal chromosomes. Let the symbol A_1 denote the event that at least one mutation has occurred on some autosomal chromosome and let A_2 denote the complementary event. It will be assumed that events A_1 and A_2 are expressed as two autosomal alleles and that A_1 is dominant to A_2. Under this assumption, there are three genotypes; namely, A_1A_1, A_1A_2 and A_2A_2. It will also be assumed that A_2 may mutate to A_1 but A_1 cannot mutate to A_2. Individuals of genotypes A_1A_1 and A_1A_2 are at risk of developing autism but individuals of genotype A_2A_2 are not at risk of developing autism. In what follows, the set of genotypes under consideration will be $\mathfrak{T} = (A_1A_1, A_1A_2, A_2A_2)$ with the genotypes numbered $\tau = 1, 2, 3$ from left to right.

By an assumption, which is based on clinical observations, females and males of genotypes A_1A_1 and A_1A_2 differ in their risk of developing autism. For the case of females with these genotypes, let π_f denote the probability that an individual develops autism, and let π_m denote a similar probability for males. In genetics, these probabilities are known as the penetrances for allele A_1 and procedures for accommodating these probabilities when computing Monte Carlo realizations of the process will be outlined subsequently. In general, penetrance is greater in males than females so that $\pi_m > \pi_f$.

As it is assumed that A_1 cannot mutate to $A_2, \mu_{12} = 0$ so that $\mu_{11} = 1$. To derive a formula of μ_{21}, the probability per generation that allele A_2 mutates to allele A_1, let μ denote the probability per generation that a mutation on some of the 44 autosomal chromosome occurs. Then, assuming that mutations on different chromosome occur independently, it follows that

$$\mu_{22} = (1 - \mu)^{44} \tag{11.9.1}$$

is the probability that no mutations occur per generation. Therefore, the probability that at least one mutation occurs on some autosomal chromosome per generation is

$$\mu_{21} = 1 - (1 - \mu)^{44} \,. \tag{11.9.2}$$

From these definitions, it follows that the mutation matrix for the model under consideration takes the form

$$\mathfrak{M} = \begin{pmatrix} 1 & 0 \\ \mu_{21} & \mu_{22} \end{pmatrix}. \tag{11.9.3}$$

Given this matrix of mutation probabilities, the gamete distributions for the three genotypes may be calculated as in table 11.5.1. Furthermore, given the matrix of mutation probabilities for genotypes, offspring probabilities of the form (11.5.2) may be calculated as well as the vector $\boldsymbol{p}(\kappa)$ in (11.5.3) for each couple type $\kappa \in \mathfrak{K}$.

Many of the components of the evolutionary model for autism may be computed by following the procedures outlined in section 11.3. In generation n, let the random functions $X(n;\tau)$ and $Y(n;\tau)$ denote the number of females and males, respectively, for each genotype $\tau \in \mathfrak{T}$ that have survived to produce offspring in generation $n+1$. Given these random functions, the random functions $N_C(n;\kappa)$ and $Z(n;\kappa)$ may be calculated as described in that section. However, in the formulation under consideration, the probability vector (11.5.5) will not be used, but attention will be focused on vectors of form $\boldsymbol{p}(\kappa)$ defined in (11.5.2) and (11.5.3) for each couple type κ. Thus, let $N(\kappa)$ denote a random variable with a Poisson distribution and parameter $\lambda(\kappa)$, let $W(\kappa;\tau)$denote a random variable representing the number of offspring of genotype τ produced by a couple of type κ per generation and let

$$\boldsymbol{W}(\kappa) = (W(\kappa;\tau) \mid \nu \in \mathfrak{T}) \tag{11.9.4}$$

denote a vector, where $\mathfrak{T}$ is the set of genotypes for both sexes. Then, given a realization of the random variable $N(\kappa)$, it will be assumed that the vector $\boldsymbol{W}(\kappa)$ has a conditional multinomial distribution with index $N(\kappa)$ and probability vector $\boldsymbol{p}(\kappa)$. In symbols,

$$\boldsymbol{W}(\kappa) \sim CMultinom(N(\kappa), \boldsymbol{p}(\kappa)). \tag{11.9.5}$$

In generation n, let the random function $T(n;\kappa,\tau)$ denote the total number of offspring of genotype $\tau \in \mathfrak{T}$ produced by couples of type $\kappa \in \mathfrak{K}$. For a given value of the random function $Z(n;\kappa) > 0$, let $W_\nu(\kappa;\tau)$ for $\nu = 1, 2, \ldots, Z(n;\kappa)$ denote a collection of conditionally independent random variables whose common distribution is defined in (11.9.5). Then, for couples of type κ, it follows that

$$T(n;\kappa,\tau) = \sum_{\nu=1}^{Z(n;\kappa)} W_\nu(\kappa;\tau), \tag{11.9.6}$$

where $T(n;\kappa,\tau) = 0$ if $Z(n;\kappa) = 0$. Let the random function $V(n;\tau)$ denote the total number of offspring of genotype τ produced by all couples in generation n. Then,

$$V(n;\tau) = \sum_{\kappa\in\mathfrak{K}} T(n;\kappa,\tau) \tag{11.9.7}$$

for all genotypes $\tau \in \mathfrak{T}$, and the total number of offspring produced in generation n is given by the sum

$$T(n) = \sum_{\tau\in\mathfrak{T}} V(n;\tau) . \tag{11.9.8}$$

Among the $V(n;\tau)$ offspring genotype τ, let $XOF(n;\tau)$ denote the number who are females. Then given $V(n;\tau)$, realizations of this random function will be computed as a sample from a binomial distribution with index $V(n;\tau)$ and probability p_f. In symbols,

$$XOF(n;\tau) \sim CBinom(V(n;\tau), p_f) . \tag{11.9.9}$$

Let the random function $YOM(n;\tau)$ denote the number of offspring of genotype τ that are males in generation n. Then, this random function is given by

$$YOM(n;\tau) = V(n;\tau) - XOF(n;\tau) \tag{11.9.10}$$

for every genotype $\tau \in \mathfrak{T}$.

Both females and males with genotypes $\tau = 1, 2$ are at risk of developing autism, and let the random function $XAUT(n;\tau)$ denote the number of females of genotypes $\tau = 1, 2$ who develop autism in generation n. Then, given a realization of $XOF(n;\tau)$, realizations of this random function will be computed as indicated in the expression

$$XAUT(n;\tau) \sim CBinom(XOF(n;\tau), \pi_f) \tag{11.9.11}$$

for $\tau = 1, 2$. Let the random function $XNAUT(n;\tau)$ denote the number of females of genotype $\tau = 1, 2$, who do not develop autism in generation n. Then, realizations of this random function would be computed according to the equation

$$XNAUT(n;\tau) = XOF(n;\tau) - XAUT(n;\tau) \tag{11.9.12}$$

for genotypes $\tau = 1, 2$. The random functions $YAUT(n;\tau)$ and $YNAUT$ $(n;\tau)$ for genotypes $\tau = 1, 2$ in generation n for males are defined similarly and realizations of these random functions are computed by using expressions similar to those in (11.9.11) and (11.9.12). In particular, the expression in (11.9.11) takes the form

$$YAUT\,(n;\tau) \sim CBinom\,(YOF\,(n;\tau)\,,\pi_m) \tag{11.9.13}$$

for males with genotypes $\tau = 1, 2$.

The next step in the description of the algorithms used to compute realizations of the stochastic model of inherited autism under consideration is to set down procedures for computing realizations of those offspring in generation n who survive to produce offspring in generation $n + 1$. Let the random function $s_f\,(n;\tau)$ denote the survival probability for females of genotype $\tau \in \mathfrak{T}$ in generation n. The parametric form of this survival probability may be chosen as in (11.6.5) for the Weibull survival function for each genotype τ. Let the random function $X\,(n + 1;\tau)$ denote the number of females in generation n of genotype τ who survive to form couples and produce offspring for generation $n + 1$. Under the assumption that only those females who develop autism will not be chosen as mates by males, it follows that for females of genotypes $\tau = 1, 2$ realizations of the random functions $X\,(n + 1;\tau)$ would be computed using the expression

$$X\,(n+1;\tau) \sim CBinom\,(XNAUT\,(n;\tau)\,, s_f\,(n;\tau))\,. \tag{11.9.14}$$

For females of genotype $\tau = 3$ however, who are not at risk of developing autism, this expression takes the form

$$X\,(n+1;\tau) \sim CBinom\,(XOF\,(n;\tau)\,, s_f\,(n;\tau)) \tag{11.9.15}$$

see (11.9.9).

Similarly for males in generation n, let the random function $Y\,(n + 1;\tau)$ denote the number of males who survive to form couples and produce offspring and let $s_m\,(n;\tau)$ denote the survival probability for genotype $\tau \in \mathfrak{T}$. Then, for genotypes $\tau = 1, 2$ realizations of these random functions are computed according to the expression

$$Y\,(n+1;\tau) \sim CBinom\,(YNAUT\,(n;\tau)\,, s_m\,(n;\tau))\,. \tag{11.9.16}$$

For, individuals of genotype $\tau = 3$ however, realizations of this random function are computed according to the expression

$$Y\,(n+1;\tau) \sim CBinom\,(YOF\,(n;\tau)\,, s_m\,(n;\tau))\,. \tag{11.9.17}$$

The last step in the description of the algorithms used to compute Monte Carlo realizations of the process under consideration is that for computing incidence of autism. The working definition of the incidence of autism in generation n is the fraction among the total number of offspring who develop autism. The total number of offspring in generation n who develop autism is given by the formula

$$TAUT\,(n) = \sum_{\tau=1}^{2} \left(XAUT\,(n;\tau) + YAUT\,(n;\tau) \right). \tag{11.9.18}$$

Therefore, the fraction of offspring in generation n who develop autism is determined by formula

$$INC\,(n) = \frac{TAUT\,(n)}{T(n)}, \tag{11.9.19}$$

see (11.9.8) for a definition of $T(n)$. Some authors define incidence as the ratio $1/INC(n)$, but this definition will not be used here, because for some realizations and generations n, the fraction may to 0 and for many programming languages $1/0$ is not defined so that if a command to compute the ratio $1/0$ is given, the execution of a program will terminate with a message of a domain error. However, when $INC\,(n) > 0$, the ratio $1/INC\,(n)$ may be computed.

Given the algorithms just outlined, one could set down procedures for computing estimates of the sample functions of the process using procedures for deriving nonlinear difference of a form similar to those displayed in section 11.7, but this derivation will be left as an exercise for an interested reader.

11.10 Autism in a Population Evolving From a Small Founder Population

In this section, the two sex model for the evolution of inherited autism described in section 11.9 will be used in a computer simulation experiment designed to study of the evolution of autism in a population evolving from a small founder population of normal females and males. For ease of reference, the two autosomal alleles under consideration will be denoted by A_1 and A_2, with A_1 dominant to A_2. The three genotypes under consideration are thus A_1A_1, A_1A_2 and A_2A_2, which will be numbered 1,2, and 3, respectively. Individuals of genotypes A_1A_1 and A_1A_2 are at risk of developing autism, but individuals of genotype A_2A_2 are normal. In experiment 11.10.1, the initial vector for females was chosen as $\boldsymbol{X}_0 = (0,0,100)$, indicating that the initial population consisted of 100 normal females of genotype A_2A_2. Similarly, it was assumed that initial population for males consisted of 105 normal individuals so that the initial vector for males was $\boldsymbol{Y}_0 = (0,0,105)$.

The number of generations of evolution considered in the experiment was chosen as $G = 2,000$ and the number of Monte Carlo replications of the experiment was $R = 100$. The time taken to complete this experiment on a 3 giga Hertz computer was about 6 hours.

Just as was the case in experiment 11.8.1, all the elements of the matrices of acceptance probabilities for females and males, $\boldsymbol{A}_f$ and $\boldsymbol{A}_m$, were chosen as 1 so that, by assumption, mating among the three genotypes with normal phenotypes was random. Furthermore, it was also assumed that the couple formation probabilities were the constant $p(\kappa) = 0.9$ for each couple type $\kappa \in \mathfrak{K}$, just as in experiment 11.8.1. At this point, however, it should be recalled that in the formulation of the model of inherited autism under consideration, females and males with symptoms of autism were excluded from the formation of couples so that, in this sense, there was selection against autism. Each element in the 3×3 matrix $\boldsymbol{\Lambda}$ for the parameters of the Poisson distributions, governing the production of offspring for each couple type, was assigned the constant value 2.5 so that there was no selection with respect to reproductive success among couple types. This value was chosen because in preliminary experiments the mutant alleles A_1 appeared in a population with high probability. Values for the rest of the parameters were chosen as in experiment 11.8.1 except that the β-parameters for both sexes were chosen as $\beta = 10^{-8}$.

The 2×2 matrix of mutation probabilities among the two alleles under consideration had the form

$$\mathfrak{M} = \begin{pmatrix} 1 & 0 \\ \mu_{21} & \mu_{22} \end{pmatrix} \tag{11.10.1}$$

where $\mu_{22} = (1-\mu)^{44}$, $\mu_{21} = 1 - (1-\mu)^{44}$ and the probability μ is to be assinged. The rationale used to make this assignment was similar to that discussed in section 5.9. That is, preliminary experiments with the embedded deterministic model were conducted for 6,000 generations of evolution by choosing several combinations of parameter values for μ, π_f and π_m and observing the incidence of autism at generation 6,000. In these experiments, it was found that the assignments $\mu = 10^{-4.001}, \pi_f = 0.49$ and $\pi_m = 0.9$ resulted in an incidence of about $INC(n) = 0.006578947$ at $n = 6,000$ generations, see (11.9.19) for a definition. Observe that this incidence leads to the value $1/INC(n) = 152$, which is close to the desired value of 150, see section 5.9 for more details. It should be noted that the value chosen for μ in this experiment was greater than the value $\mu = 10^{-6}$ used in the experiments reported in section 5.9. From this observation, one may conclude that, given a fixed values of incidence, the value of μ chosen to attain this incidence depends not only on the model under consideration but also on the values chosen for the other parameters of a model. This ambiguous situation will often arise when there are insufficient data to estimate the mutation probability μ as was the case for the model of inherited autism considered in this section as well as in section 5.9.

Presented in Figure 11.10.1 are the graphs of the MAX, MIN, and $Q50$ trajectories for the incidence of autism as estimated from a sample Monte Carlo realizations of the stochastic process, governing the evolution of autism in the hypothetical population under consideration, for the first 1,000 of the 2,000 generations in the experiment. To provide a basis for comparing of the trajectory of the incidence of autism computed by using the embedded deterministic with the estimated trajectories of the stochastic shown in the figure, the deterministic trajectory DET was also plotted. As can be seen from the figure, at about 150 to 175 generations into the projection, the four trajectories under consideration had reached values that were so close that they could not be distinguished on the scale of the vertical axis of the graphs. However, before the four trajectories leveled off, the MAX and the MIN trajectories exhibited relatively high levels of fluctuations. As indicated above, the reciprocals of the trajectories after 200 generations were about 152.

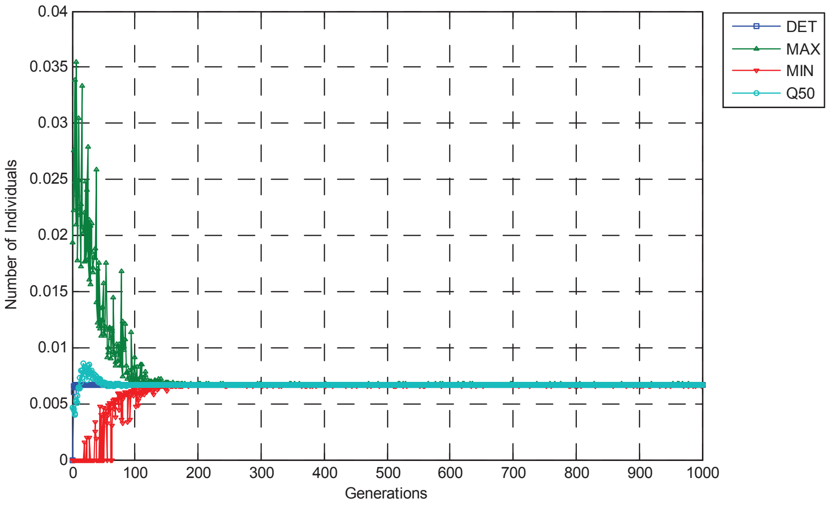

Figure 11.10.1 The Evolution of the Incidence of Autism.

It of interest to compare Figure 11.10.1 with Figures 5.9.1 and 5.9.2, where the incidence of autism, using the model for the evolution of hereditary autism described in section 5.4, was plotted. As can be seen by comparing these figures, convergence to a constant for the deterministic model and to a quasi-stationary distribution for the stochastic model was much more rapid as displayed in figure 11.10.1 than in the gamete sampling model described in chapter 5. For from an inspection of figure 5.9.1, it can be seen that there was no suggestion of convergence until nearly 400 generations into the projection; whereas in Figure 11.10.1, it can be seen that convergence had occurred somewhere between about 150 and 175 generations. It should be emphasized, however, that the formulations underlying the genetic models of hereditary autism considered in this chapter and Chapter 5 are fundamentally different. For in the model considered in Chapter 5, it was assumed that population size was constant from generation to generation, but in the self regulatory branching process under consideration in this section, population size varied from generation to generation and it was supposed that the population evolved from a small founder population.

The motivation of the graphs of the trajectories for the three genotypes under consideration presented in Figure 11.10.2 was to obtain an overview

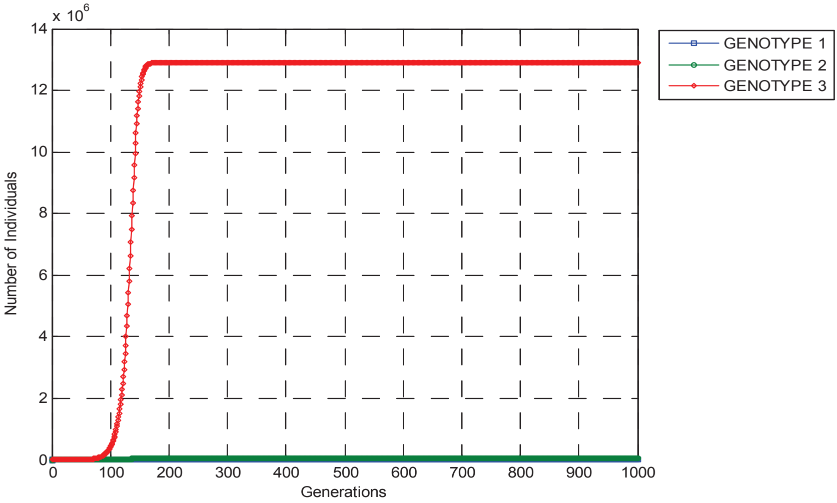

Figure 11.10.2 Estimated Deterministic Trajectories of the Number of Individual in the Female Population of the The Three Genotypes in Experiment 11.10.1.

of the evolution of the numbers of individuals of each of the genotypes A_1A_1, A_1A_2 and A_2A_2 in the female population during the first 1000 generations of evolution as estimated from the embedded deterministic model. As can be seen from the figure, at about 150 generations into the projection, the number of individuals of the normal genotype 3 had reached a plateau of about 13×10^6 individuals, where the numbers of individuals of genotypes 1 and 2, who were at risk of developing autism, remained small with respect to the scale used on the vertical axis. But, as will be shown in the next figure, the number of individuals of genotype 2, A_1A_2, for the female population had actually reached a number of about 6.5×10^4 individuals.

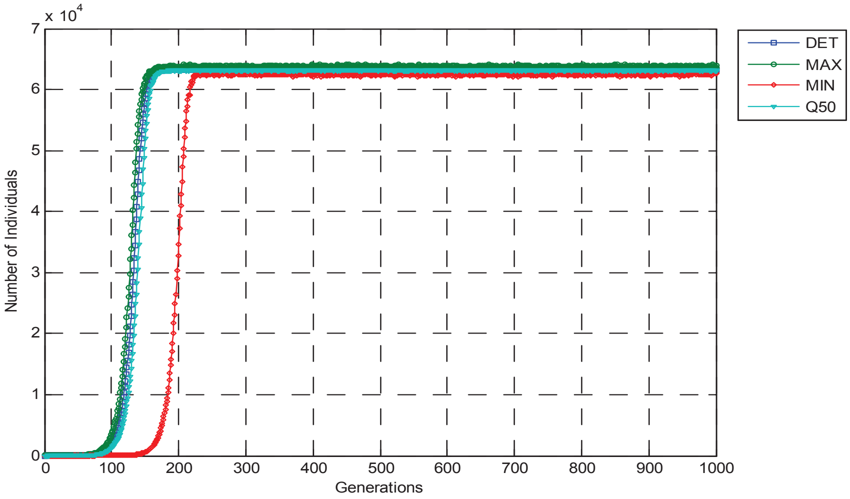

Figure 11.10.3 Estimated DET, MAX, MIN and $Q50$ Trajectories for the Heterozygous Genotype A_1A_2 in the Female Population.

To show that the number of individuals in the female population with heterozygous genotype 2 had reached a plateau of about 6.5×10^4 individuals, in Figure 11.10.3 the DET, MAX, MIN and $Q50$ trajectories are plotted for the first 1000 generations of the projection. From this figure, it can be seen that the DET, MAX and $Q50$ are very close on the vertical scale of the figure as they rise to a level of about 6.5×10^4 individuals at a little more than 100 generations into the projection. The MIN trajectory, however, rises much more slowly and does not reach levels of about 6.5×10^4 individuals until about 200 generations into the projection. This slower rise of the MIN trajectory is indicative of the relatively high levels of variation among the realizations of the stochastic process during the first 200 generations of the projection. The graphs presented in figure 11.10.3 are significant, because they supply information about how many heterozygous females are present in the population out of a total about 13×10^6 females who are significant contributors to the rather high incidence of autism due to the lower penetrance of the mutant allele A_1 for females.

11.11 Sexual Selection in Populations Evolving From a Small Founder Population

As early as 1871, Darwin (1871) in his book described two mechanisms by which natural selection could occur as a result of the interactions of individuals within and between sexes. One of these mechanisms was the physical struggles that occur when males compete for females. In such cases, males who dominate in such struggles would be the ones most likely to mate with females. A second mechanism in which selection could occur is that in which females prefer males with certain phenotypes as mates. A reader who wishes to delve into the subject of sexual selection in greater depth may consult the book by Hartl and Clark (1989) for a discussion of literature on the subject and further examples of sexual selection. In this section, attention will be confined primarily to two computer simulation experiments in which it is supposed that females prefer males with a mutant genotype as sexual partners. Furthermore, as in other computer experiments reported in this and other chapters, it will be supposed that the population evolves from a small founder population. A third experiment with the embedded deterministic model will also be briefly discussed in which it was assumed the population evolved from a population in demographic equilibrium which was punctuated by the occurrences of new mutations.

In experiment 11.11.1, an autosomal locus with two alleles A and a was under consideration in which it was assumed that allele A was dominant to allele a. The three genotypes under consideration will be denoted by AA, Aa and aa and will be numbered 1,2,3. Moreover, the initial vector for females was chosen as $\boldsymbol{X}_0 = (100, 0, 0)$ and that for males was chosen as $\boldsymbol{Y}_0 = (105, 0, 0)$ as in experiment 11.8.1, and the matrix of mutation probabilities was chosen as in (11.8.1). With the exception of the matrix of acceptance probabilities for females $\boldsymbol{A}_f$, the 3×3 matrix $\boldsymbol{\Lambda}$ of Poisson parameters for the nine types of matings and the number of generations $G = 3,000$ considered, all parameter values used in experiment 11.11.1 were the same as those for experiment 11.8.1, see section 11.8. To take into account the assumption that females of every genotype preferred males of genotype aa as mates, the elements in the matrix of acceptance probabilities for females were assigned the values

$$\boldsymbol{A}_f = \begin{pmatrix} 0.1 & 0.1 & 0.9 \\ 0.1 & 0.1 & 0.9 \\ 0.1 & 0.1 & 0.9 \end{pmatrix}. \tag{11.11.1}$$

However, it was assumed that males of the three genotypes mated at random with females, which was characterized by letting all elements in the 3×3 matrix of $\boldsymbol{A}_m$ acceptance probabilities for males equal one. It was also assumed that there was no differential selection with respect to the expected number of offspring produced by each mating type and it was supposed that all elements in the matrix $\mathbf{\Lambda}$ were equal to 2.5. The time taken to compute a sample, consisting of $G = 3,000$ generations with $R = 100$ replications, of Monte Carlo realizations of the process was about 17 hours on a desktop computer that operated at speed of 3 giga Hertz.

Figure 11.11.1 contains the graphs of the trajectories of the three genotypes for the female population, which were computed by using the deterministic model embedded in the stochastic process. As can be seen from this figure, starting from an initial number of 100 females of genotype AA, the graph of the trajectory for genotype 1 rises steeply and reaches a maximum somewhere in the neighborhood of 50 to 100 generations. It then starts a slow decline which begins to accelerate dramatically between 2000 and 2500 generations into the projection and then reaches a low point shortly after 2500 generations. The number of individuals of the heterozygote Aa, increases slowly during the first 1000 generations of the projection, and then increases more rapidly during the time span of 1000 to 2500 generations but declines sharply after generation 2500, see the trajectory of genotype 2. As can be seen from the trajectory of genotype 3, throughout all generations from 0 to about 2500, the number of individuals of the mutant recessive genotype aa remained very small, but after this number reached a threshold sometime after generation 2500, the trajectory of genotype 3 rises sharply and became the predominant genotype in the population somewhere between 2500 and 3000 generations. It is of interest to note by observing the figure that the rise to predominance of genotype 3 did not occur until last the 500 generations of a projection consisting of 3000 generations. Interestingly, the graphs of deterministic trajectories for the three genotypes for the male population are very similar to those for females and were, therefore, omitted.

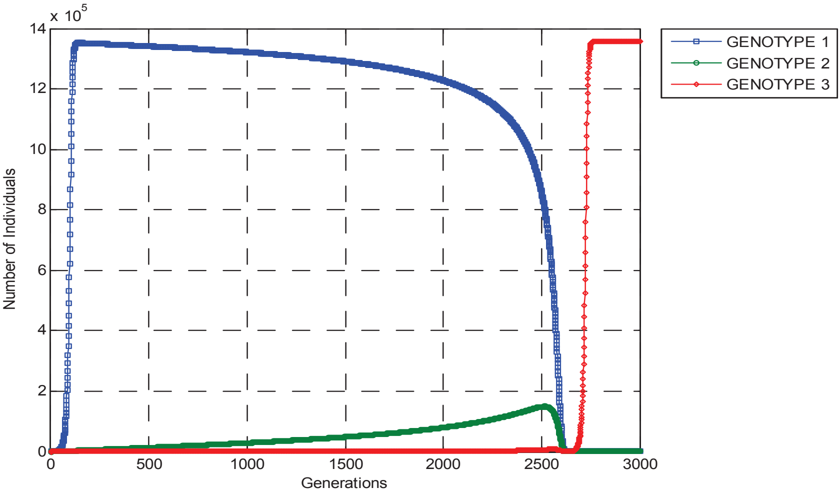

Figure 11.11.1 Deterministic Trajectories of the Three Genotypes for the Female Population in Experiment 11.11.1.

Presented in Figure 11.11.2 are the $MAX, MEAN, Q50$ and SD trajectories for females of the recessive genotype 3, aa, as statistically summarized from the sample of 100 Monte Carlo realizations of the process. Because individuals of genotype 3 did not appear in the population in significant numbers until 2500 generations into stochastic projections of 3000 generations, these trajectories were plotted only for generations 2500 to 3000. As can be seen from this figure, the MAX trajectory for genotype 3 rose sharply after generation 2700 of the projection and reached a maximum of somewhat less than 14×10^5 individuals shortly thereafter. It is of interest to note that the MAX trajectory closely resembles that of the embedded deterministic model. On the other hand, throughout the 500 generations of the stochastic projection shown in the figure, the $Q50$ remained small, indicating that in 50 of the replications of the process there were relatively few individuals of the recessive genotype aa. This wide spread between the MAX and $Q50$ trajectories is indicative of a high level of variability, stochasticity, among the realizations of the process. A glimpse of the high level of stochasticity may also be obtained by comparing the $MEAN$ and standard deviation trajectories, SD, in the figure. For, as can be seen from the figure, after generation 2700 of the projection, the SD trajectory is uniformly greater than the $MEAN$ trajectory, which is a signature of a high level of stochasticity among the realizations of the process. The corresponding trajectories for males of genotype 3 were similar to those for females and, therefore, graphs of these trajectories will not be shown.

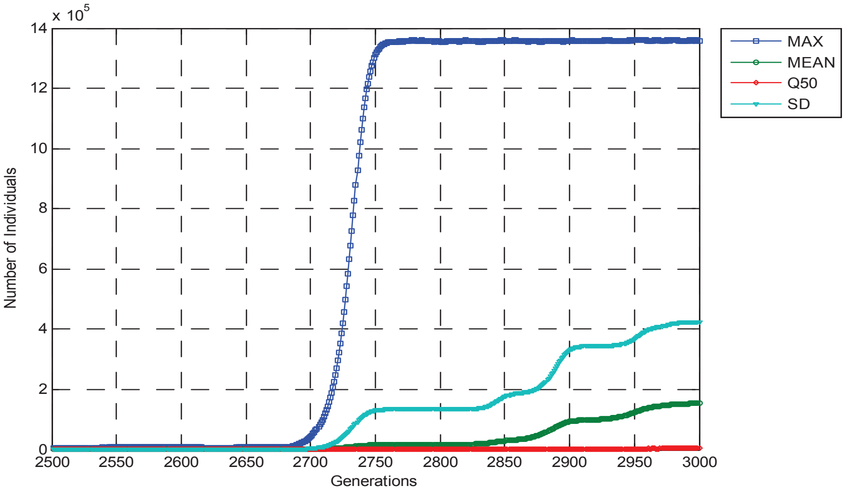

Figure 11.11.2 Graphs of the MAX, $MEAN$, $Q50$ and SD Trajectories of the Process for Females of Genotype 3 as Estimated From a Monte Carlo Sample for Experiment 11.11.1.

In experiment 11.11.2, all the parameter assignments used in experiment 11.11.1 were used with the exception of the parameters in the matrix $\mathbf{\Lambda}$, which were chosen as the constant value 4. Thus, in this experiment it was assumed that each couple type on average had four offspring. The reason for choosing this rather high value for the expected number of offspring per couple was that it was desired to investigate a case in which evolution would be more rapid and whether, with this more rapid evolution, the level of stochasticity among the realizations of the process would be less than in experiment 11.11.1. Presented in Figure 11.11.3 are the graphs of the trajectories of each of the three genotypes as computed using the embedded deterministic model. As can be seen from an inspection of this figure, the pace of evolution was somewhat more rapid than was the case in experiment 11.11.1, where all couples produced on average 2.5 offspring. For in this experiment, the trajectory for genotype 1 reached its maximum value of over 2×10^6 individuals at fewer generations into the projection than in experiment 11.11.1, see Figure 11.11.1. On the other hand, the overall pattern of the graphs of the three genotypes in Figure 11.11.3 are quite similar to those in Figure 11.11.1 in that in both experiments, the steep rise in the number of individuals of genotypes 3 occurred only after 2500 generations into the projection.

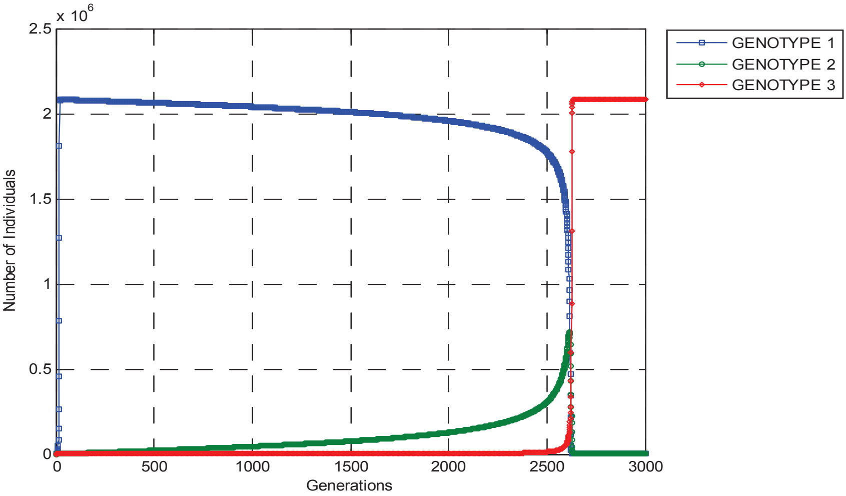

Figure 11.11.3 Deterministic Trajectories of the Three Genotype in the Female Population for Experiment 11.11.2.

Contained in Figure 11.11.4 are the graphs of selected trajectories for females of genotype 3, which were estimated from the 100 Monte Carlo realizations of the experiment. From this figure, it can be seen that the pace of evolution in this experiment was more rapid than in experiment 11.11.1. For in experiment 11.11.2, the MAX trajectory for the recessive genotype aa rose to a level of over 2×10^6 individuals before 2600 generations into the projection of 3000 generations. Furthermore, the MIN trajectory was positive for most of the generations shown in the figure in contrast with a flat zero MIN trajectory, which was not shown in Figure 11.11.2. Unlike the $Q50$ trajectory displayed in Figure 11.11.2, this trajectory for genotype 3 in this experiment underwent a steep rise to over 2×10^6 individuals on or about 2600 generations into the projection. At about this generation into the projection the $MEAN$ trajectory crosses the standard deviation trajectory SD and at 3000 generations the mean trajectory actual exceed the SD, which is an indicator for a lower level of stochasticity in experiment 11.11.2 than in experiment 11.11.1, see Figure 11.11.2.

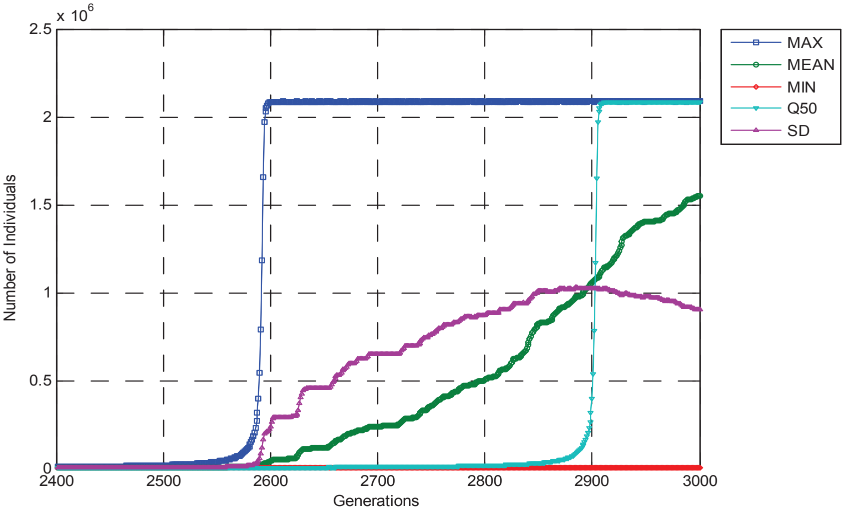

Figure 11.11.4 Selected Trajectories for the Stochastic Model for Females of Genotype 3 in Experiment 11.11.2.

In experiment 11.11.3 all parameter assignments were the same as those in experiment 11.11.2 except that a different set of initial conditions were used. For in this experiment, the initial vector for the female population was chosen as $\boldsymbol{X}_0 = (2000000, 0, 0)$, indicating that were 2,000,000 females of genotype AA. The initial vector for the male population was chosen as $\boldsymbol{Y}_0 = 1.05 \times \boldsymbol{X}_0$, which resulted in the vector $\boldsymbol{Y}_0 = (2100000, 0, 0)$, indicating there were 2,100,000 males of genotype AA. In a preliminary experiment with the self regulating branching process under consideration, it was observed that these numbers were close to those observed when the embedded deterministic model converged to a limit or equilibrium when the parameters used in experiment 11.11.3 were in force. Therefore, in this experiment the founder population was homozygous for genotype AA and was, for all practical purposes, in a demographic equilibrium, and, metaphorically speaking, this equilibrium was punctuated by introducing positive probabilities for mutation into the experiment. The number of generations of evolution considered using the embedded deterministic model was $G = 6,000$, and the time taken to execute this experiment was a little over 7 seconds.

This execution time is very similar to those observed in other experiments with the deterministic model, which is a desirable property when one wishes to quickly explore the implications of an assinged set of parameter values. It should be emphasized, however, that if one confines attention to the deterministic model in conducting computer experiments, the rather high levels of stochasticity that are usually observed in the initial stages of a projection before the process converges to a quasi-stationary distribution would be missed. As expected, according the projection based on the deterministic model, mutant genotypes appeared much more rapidly than in experiments 11.11.1 and 11.11.2, where the population evolved from a small founder population. However, the rise of the predominance of genotypes aa in the populations was relatively slow, because the rapid convergence to a limit was observed somewhere between generations 2,600 to 2,700, which was somewhat more rapid than that observed in experiments 11.11.1 and 11.11.2. As the trajectory for genotype aa was very similar to those shown in experiments 11.11.1 and 11.11.2 as shown in Figures 11.11.1 and 11.11.3, the graphs of the trajectories of the three genotypes in the female population will be omitted.

11.12 Two Sex Processes with Linkage at Two Autosomal Loci

In this section, some of the issues will be discussed that arise when attempts are made to extend the two sex processes treated in the foregoing sections of this chapter to cases of two linked autosomal loci. Suppose each locus has two alleles and let A and a denote the two alleles at locus one, and, similarly, let B and b denote the two alleles at locus two. Following the notation that was used in Chapters 2 and 3, let the symbol (x_1x_2, y_1y_2) denote a general genotype of some individual, where the symbol x_1x_2 denotes the two alleles contributed by the female parent and the symbol y_1y_2 denotes the two alleles contributed by the male parent. To help fix ideas, the symbol x_1 denotes either allele A or a and the symbol x_2 denotes either allele B or b. The pair of symbols y_1y_2 will be interpreted similarly. Observe that among the four symbols which are used to denote a general genotype (x_1x_2, y_1y_2)

each symbol may be one of two forms. Therefore, the set $\mathfrak{G}$ of all genotypes with respect to the two loci under consideration with 2 alleles at each locus will contain 16 distinct genotypes when the contributions of the female and male parents of each individual are taken into account.

A general genotype of the form (x_1x_2, y_1y_2) may produce four types of gametes as a result of a type of reduction cell division called meiosis. In the general case, these four types of gametes will be denoted by x_1x_2, and y_1y_2, x_1y_2 and y_1x_2. The gamete x_1x_2 represents the original maternal alleles at the two loci under consideration and the gamete y_1y_2 represents the original combinations of the paternal alleles at the two loci. The gametes x_1y_2 and y_1x_2 however, represent recombination events, because these gametes contain combinations of maternal and paternal alleles. Let the symbol $\gamma\,(0,0)$ denote the probability of finding a gamete of type x_1x_2 in the gamete pool of an individual with genotype (x_1x_2, y_1y_2) and let $\gamma\,(1,1)$ denote the probability of finding a gamete of type y_1y_2 in the gamete pool of this individual. As in Chapters 2 and 3, the symbol 0 denotes that an allele is a copy of a maternal allele and the symbol 1 denotes the allele is a copy of a paternal allele. Similarly, let the symbols $\gamma\,(0,1)$ and $\gamma\,(1,0)$ denote the probabilities of finding recombinant gametes x_1y_2 and y_1x_2, respectively, in the gamete pool of this individual. Next let ρ denote the recombination probability for the two loci under consideration. Then, because of the symmetry of the meiotic process, $\gamma\,(0,0) = \gamma\,(1,1) = (1-\rho)\,/2$ and $\gamma\,(0,1) = \gamma\,(1,0) = \rho/2$. Given this parameterization of linkage for the two autosomal loci under consideration, it will be possible to derive the gametic distribution for each of the 16 genotypes under consideration. Chapters 2 and 3 may be consulted for more detailed definitions of the notation being used in this section. For the sake of simplicity in what follows, it will also be assumed that there is no mutation at each of the two loci under consideration.

As a first step in setting up a procedure for the derivation of the gametic distribution for each genotype, the four types of gametes under consideration with be represented in the order AB, Ab, aB, ab. By definition, the gametic distribution for any genotype (x_1x_2, y_1y_2) is the set of four nonnegative numbers $(g\,(AB)\,, g\,(Ab)\,, g\,(aB)\,, g\,(ab))$, whose sum is one. Given this definition and ordering of gamete types, consider the gametic distribution of the homozygous genotype $(AB, AB) = (x_1x_2, y_1y_2)$. There are four distinct gametes for such individuals that consist of various combinations of maternal and paternal alleles. In general case, the set of four gametes for the genotype under consideration is

$$(x_1x_2, x_1y_2, y_1x_2, y_1y_2) = (AB, AB, AB, AB)\,, \tag{11.12.1}$$

which occur with probabilities $\gamma(0,0), \gamma(0,1), \gamma(1,0), \gamma(1,1)$, respectively. But, because of the homozygosity of the individual under consideration, these four combinations of maternal and paternal alleles are all the same type of gamete; namely, AB as shown in the equation (11.12.1). Hence, for this homozygous genotype, $g(AB) = \gamma(0,0) + \gamma(0,1) + \gamma(1,0) + \gamma(1,1) = 1$ so that the gametic distribution for the genotype (AB, AB) is $(1,0,0,0)$. Of course, this result is intuitively obvious when there is no mutation, but, nevertheless, the derivation just described is informative, particularly if one wished to take mutation into account at the two loci.

Next consider the genotype $(AB, Ab) = (x_1x_2, y_1y_2)$, which is homozygous at locus 1 but heterozygous at locus 2. In this case, the set of four combinations of maternal and paternal alleles has the form

$$(x_1x_2, x_1y_2, y_1x_2, y_1y_2) = (AB, Ab, AB, Ab), \tag{11.12.2}$$

which occur with probabilities $\gamma(0,0), \gamma(0,1), \gamma(1,0), \gamma(1,1)$, respectively. Observe that $\gamma(0,0) + \gamma(1,0) = (1-\rho)/2 + \rho/2 = 1/2$ is the probability of observing a gamete of type AB in the gamete pool of this genotype. Similarly, the probability of observing a gamete of type Ab in this pool is $1/2$. Therefore, the gametic distribution for this genotype is $(1/2, 1/2, 0, 0)$.

By applying this procedure, it is possible to derive the gametic distribution for each of the 16 genotypes under consideration. Presented in the table 11.12.1 are the gametic distributions for the set of four genotypes such that the gamete type AB was contributed by the female parent.

Table 11.12.1 Gametic Distributions for the Set of Genotypes Such That Gamete Type ***AB*** Was Contributed by the Female Parent

○	***AB***	***Ab***	***aB***	***ab***
(AB, AB)	1	0	0	0
(AB, Ab)	$\frac{1}{2}$	$\frac{1}{2}$	0	0
(AB, aB)	$\frac{1}{2}$	0	$\frac{1}{2}$	0
(AB, ab)	$\gamma(0,0)$	$\gamma(0,1)$	$\gamma(1,0)$	$\gamma(1,1)$

Table 11.12.2 contains the set of gametic distributions for the set of four genotypes such that the maternal contribution was the gamete type Ab.

For the sake of completeness, the gametic distribution for each of the eight remaining genotypes will also be listed. In the table 11.12.3 is a list of gametic distribution for each of those four genotypes such the that the gamete type contributed by the female was aB.

Table 11.12.2 Gametic Distributions for the Set of Genotypes Such That Gamete Type ***Ab*** Was Contributed by the Female Parent

∘	***AB***	***Ab***	***aB***	***ab***
(Ab, AB)	$\frac{1}{2}$	$\frac{1}{2}$	0	0
(Ab, Ab)	0	1	0	0
(Ab, aB)	$\gamma\,(0,1)$	$\gamma\,(0,0)$	$\gamma\,(1,1)$	$\gamma\,(1,0)$
(Ab, ab)	0	$\frac{1}{2}$	0	$\frac{1}{2}$

Table 11.12.3 Gametic Distributions for the Set of Genotypes Such That Gamete Type ***aB*** Was Contributed by the Female Parent

∘	***AB***	***Ab***	***aB***	***ab***
(aB, AB)	$\frac{1}{2}$	0	$\frac{1}{2}$	0
(aB, Ab)	$\gamma\,(1,0)$	$\gamma\,(1,1)$	$\gamma\,(0,0)$	$\gamma\,(0,1)$
(aB, aB)	0	0	1	0
(aB, ab)	0	0	$\frac{1}{2}$	$\frac{1}{2}$

Finally, contained in the table 11.12.4, are the gametic distribution for the four genotypes such that the type of gamete contributed by the females was ab.

Table 11.12.4 Gametic Distributions for the Set of Genotypes Such That Gamete Type ***ab*** Was Contributed by the Female Parent

∘	***AB***	***Ab***	***aB***	***ab***
(ab, AB)	$\gamma\,(1,1)$	$\gamma\,(1,0)$	$\gamma\,(0,1)$	$\gamma\,(0,0)$
(ab, Ab)	0	$\frac{1}{2}$	0	$\frac{1}{2}$
(ab, aB)	0	0	$\frac{1}{2}$	$\frac{1}{2}$
(ab, ab)	0	0	0	1

It would be of interest to include the possibility of mutation among the gametes in the two loci model under consideration, but, in this section, only some of the issues that arise when attempts are made to incorporate mutation into the formulation will be discussed. Let the indices $i = 1, 2, 3, 4$ denote, respectively, the four types of gametes under consideration; namely AB, Ab, aB and ab. Moreover, let

$$\mathfrak{M} = \left(\mu_{ij}\right) \tag{11.12.3}$$

denote a 4×4 matrix of mutation probabilities. In this matrix $\mu_{ij} \geq 0$ is the conditional probability that a gamete of type i mutates to a gamete of type $j \neq i$ per generation, and μ_{ii} is the conditional probability that a gamete of type i does not mutate per generation. Like all other mutation matrices considered in the foregoing sections of this chapter, each row of the matrix $\mathfrak{M}$ sums to one. The set of gametic distribution enumerated above for the case of no mutation, will also play a fundamental role when mutation among the types of gametes is accommodated in the model. Let $\mathfrak{T}$ be the set of 16 genotypes under consideration, and for each $\tau \in \mathfrak{T}$, let the 1×4 vector $\boldsymbol{g}(\tau)$ denote the gametic distribution when there is no mutation. But, when mutation among the gametes is taken into account, this vector will need to be adjusted. Let $\boldsymbol{g}_{mut}(\tau)$ denote the gametic distribution of genotype τ adjusted for mutation among the four types of gametes under consideration. Then, this adjusted distribution may be determined by the formula

$$\boldsymbol{g}_{mut}(\tau) = \boldsymbol{g}(\tau)\,\mathfrak{M}. \tag{11.12.4}$$

Formula $(11, 12, 4)$ embodies the notion that the vector $\boldsymbol{g}(\tau)$ accounts for recombination events that occur prior to mutation, and, given these recombination events, the matrix $\mathfrak{M}$ contains the conditional probabilities of mutation among gametes per generation. Actually, little seems to be known about the interaction of recombination and mutation during process of meiosis, but equation (11.12.4) is a useful approach to formulating a model of genetic recombination and mutation that has desirable mathematical properties. For example, if $\mathfrak{M} = \boldsymbol{I}$, the 4×4 identity matrix, which implies there is no mutation among the four types of gametes, then $\boldsymbol{g}_{mut}(\tau) = \boldsymbol{g}(\tau)$ as it should. Observe that in any computer implementation of model under consideration, formula (11.12.4) would need to be evaluated for each of the 16 genotypes $\tau \in \mathfrak{T}$, which would entail considerable computation.

Another set of issues arise, when one considers extending the two sex models discussed in the foregoing sections of this chapter in which only two alleles at some autosomal locus were under consideration, to cases in which two linked loci are incorporated into the formulation. Observe that when 16 genotypes are under consideration, then the set $\mathfrak{K}$ of couple types contains $16 \times 16 = 256$ elements. Therefore, the matrix of acceptance probabilities for females $\boldsymbol{A}_f$ is of order 16×16 and contains 256 parameters. Similar remarks hold for the matrix $\boldsymbol{A}_m$ of acceptance probabilities for males. When it is assumed that the mating is random among the 16 genotypes, then assigning values to the 256 parameters in each of these matrices would be straight forward, because each element in these matrices could be assinged the number 1. On the other hand, if one wished to accommodate a form of natural selection based on differential expectations of offspring produced by each couple type under random mating, one would need to consider ways to simplify the number of phenotypes that would need to be considered. For in this case, the matrix $\mathbf{\Lambda}$ of the expectations of the number of offspring for each couple type would also be of order 16×16 and contain 256 parameters.

One approach to reducing the number of parameters that need to be considered arises when it is assumed that the alleles A and B are both dominant to their respective recessive alleles a and b. Under this assumption, there are only four phenotypes to consider, which may be symbolized as $A - B-, A - bb, aa, B-$ and $aabb$. Therefore, under the assumption that mates for females and males are selected on the basis of phenotype, it follows that in each row of the acceptance matrices $\boldsymbol{A}_f$ and $\boldsymbol{A}_m$ for females and males would contain at most four distinct parameters, which would greatly simplify the problem of assigning parameter values in each of these matrices. Furthermore, if it is assumed that the expected number of offspring produced by each couple type $\kappa \in \mathfrak{K}$ depends on only the phenotypes of the female and male of the couple, then the number of distinct parameters in the 16×16 matrix $\mathbf{\Lambda}$ would be reduced to a workable number.

Thus, for cases in which the dynamics of the evolutionary process are determined by the phenotypes of individuals, it would be feasible to study the evolution of such populations in computer simulation experiments. In such experiments, two components of natural selection could be studied; namely a component based on the selection of mates for both females and males and a component based on reproductive success for each couple type. In addition, if it is desired to formulate a self regulating stochastic process, then by assuming the competitive ability of an individual depends only on its phenotype, it would suffice to consider two parameters in each of the four Weibull survival functions for each sex.

In conclusion, given the above list of gametic distributions for the 16 genotypes, it would be a straight forward exercise to write software to implement the case of no selection and mutation with random mating that included survival probabilities for each of the 16 genotypes for both sexes to control eventual population size. Such software would be of interest, because it would provide a means for studying the rate of convergence to linkage equilibrium in a finite population, which in turn would be useful tool in genome wide searches for signatures of natural selection in which linkage disequilibrium is used as an indicator of regions of the genome that may be implicated in selection, natural or otherwise.

Writing software for formulations that include components of natural selection as well as non-random mating would be much more challenging.

Nevertheless, efforts to include such factors in the model should be encouraged. Another benefit that may be derived from such software is that providing a capability for studying selective sweeps under non-constant population size when one allele at one locus has a selective advantage and the alleles at the second linked locus are neutral. For more technical details on selection sweeps, the paper by Durrett and Schweinsberg (2004) may be consulted as well as the references cited in that paper.

Bibliography

[1] Darwin, C. (1871) **The Descent of Man and Selection in Relation to Sex**. D. Appleton and Co., New York.

[2] Durrett, R. and Schweinsberg, J. (2004) Approximating Selective Sweeps. **Theoretical Population Biology 66**:129–138.

[3] Durrett, R. (2008) **Probability Models for DNA Sequence Evolution, Second Edition**. Springer Science.

[4] Hartl, D. L. and Clark, A. G. (1989) **Principles of Population Genetics, Second Edition**. Sinauer Associates, Inc., Sunderlan, Massachusetts.

[5] Mode, C. J. (1995) An Extension of the Galton-Watson Process to a Two Sex Density Dependent Genetic Model. **Proceeding of the First World Conference on Branching Processes**. Edited by C. Heyde, Springer Lecture Notes in Statistics. **99:**152–168.

[6] Mode, C. J. and Sleeman, C. K. (2000) **Stochastic Processes in Epidemiology - HIV/AIDS, Other Infectious Diseases and Computers**. World Scientific, New Jersey, London, Hong Kong.

[7] Nowak, M. A. (2006) **Evolutionary Dynamics - Exploring the Equations of Life**. The Belknap Press of Harvard University Press. Cambridge, Mass. and London, England.

Chapter 12

Multitype Self-Regulatory Branching Process and the Evolutionary Genetics of Age Structured Two Sex Populations

12.1 Introduction

In this chapter, two sex age dependent stochastic processes with couple formation are formulated with a view towards developing software to implement them on desk top computers. Just as in the foregoing chapters of this book, attention was also devoted to embedding deterministic models within a stochastic process. When there are fifty or more age classes under consideration in a formulation and it is also assumed that more than one diploid genotype may be present at any time, then the number of couple types may become very large, which necessitates the processing of large arrays repeatedly in a computer for each time unit in an evolutionary experiment that will generally include thousands of time units such as years.

Processing large arrays becomes particularly problematic on desk top computers with limited memory. One approach to reducing the size of the arrays that need to be processed in a computer is to ignore the formation of couples and confine attention to an age dependent formulation with two sexes with one or more genotypes in which sexual contacts occur among females and males. Sexual contacts may entail randomness with respect to ages and genotypes for both females and males or some assortative schemes for choosing sexual partners according to genotypes and ages may also be introduced into a formulation. A second approach was to aim for the computer implementation of a process with couple formation with a rather small number of age classes that will reduce the size of arrays that must be processed in a computer. Software was written and implemented for both these approaches, but in all cases in preliminary computer experiments designed to explore some of the properties of each formulation, attention was focused only on the embedded deterministic model so as to gain some experience with the times needed to complete computer experiments involving thousands of time units of evolution.

The importance of age dependent models in describing the evolution of age structured populations, such as humans, other mammalian species and various species of birds, has been recognized by many authors. For more information on age structured models it is suggested that the book by Charlesworth (1980) be consulted as well as the references contained therein. Unlike the models which are formulated in this chapter, however, most of the discussion in the book by Charlesworth is centered on deterministic models. In the field of human demography, age dependent models have also been used extensively, but, for the most part, attention has been confined to deterministic models.

An extensive use of stochastic models in human demography, based of generalized branching process in discrete time, were used by Mode and were summarized in the book Mode (1985). Contained in this book are many references to the literature in human demography. In particular, much attention was devoted to stochastic models of human reproduction which accommodate many features of the reproductive physiology of the human female, which were applied in evaluating the impact of family planning programs on human fertility. Because these models of human reproduction have a rather complicated structure, they would over burden already rather complicated formulations to be described, and will, therefore, not be included in this chapter.

With regard to the literature on models used in the study of the evolution of age structured populations, the models described and applied in this chapter seem novel in the sense that attention was confined solely to stochastic processes that belong to a class of self regulating multitype branching processes in discrete time, even though the writing of software to implement such models has been delayed until desk top computers with more memory and higher speeds of execution become available.

Another feature that seems to be novel in the study of the evolution of age structured populations is that of systematically embedding deterministic models in stochastic processes that may implemented on a computer with relative ease and with the property that computer experiments involving thousands of time units of evolution may be completed within relatively short computer execution times. Of course the writing of software by the authors was not novel, but it was, nevertheless, important, because it clearly demonstrates that by so doing interesting insights into the evolution of age structured populations may be gained from computer experimentation, particularly in the emergence of new beneficial mutations.

12.2 An Overview of Competing Risks and Semi-Markov Processes

When considering the construction of age-dependent evolutionary processes, a tool that is very useful when formulating the process is that of a risk function. By way of defining this function, let T denote a random variable with support $[0, \infty) = \mathbb{T}^{+}$, which may be interpreted as the waiting time to some event, and for all $t \in \mathbb{T}^{+}$, let $F(t)$ denote its distribution function

$$F(t) = P[T \leq t]. \tag{12.2.1}$$

Let $f(t)$ denote a continuous probability density function, *p.d.f.*, of the random variable T. By definition the survival function is

$$S(t) = P(T > t) = 1 - F(t) = \int_{t}^{\infty} f(s)\, ds, \tag{12.2.2}$$

which is defined for all $t \in \mathbb{T}^{+}$. A formal definition of the risk function $\theta(t)$ for the distribution under consideration for $t \in \mathbb{T}^{+}$ is

$$-\frac{d \ln S(t)}{dt} = \theta(t) = \frac{f(t)}{S(t)}. \tag{12.2.3}$$

From this equation, it can be seen that $\theta(t)$ may be defined in terms of the *p.d.f.* and the survival function. Moreover, for $h > 0$, it may also be interpreted in terms of the conditional probability

$$\frac{P[t < T \leq t + h]}{P[T > t]} \simeq \theta(t)\, h, \tag{12.2.4}$$

where the symbol $\simeq$ denotes an approximation. Observe that this is the conditional probability that the event occurs in the time interval $(t, t+h]$, given that it occurs sometime after t. From equation (12.2.3) it can also be seen that if the risk function $\theta(t)$ is specified in advance, then it also determines the distribution function. For, if it is assumed that $\theta(t)$ in continuous on $\mathbb{T}^{+}$, then by integrating (12.2.3) and using the initial condition $S(0) = 1$, it follows that

$$S(t) = \exp\left(-\int_{0}^{t} \theta(s)\, ds\right) \tag{12.2.5}$$

for all $t \in \mathbb{T}^+$. Therefore,

$$F(t) = 1 - S(t) = 1 - \exp\left(-\int_0^t \theta(s)\,ds\right) \tag{12.2.6}$$

and the *p.d.f.* of the random variable T is

$$F'(t) = f(t) = \theta(t)\exp\left(-\int_0^t \theta(s)\,ds\right) \tag{12.2.7}$$

for all $t \in \mathbb{T}^+$.

By way of a simple example, suppose that the risk function of distribution is constant $\beta > 0$ so that $\theta(t) = \beta$ for all $t \in \mathbb{T}^+$. Under this assumption, it can be seen that the integral on the right in (12.2.7) reduces to βt. Therefore, the *p.d.f.* of a distribution with a constant risk function is

$$f(t) = \beta\exp(-\beta t) \tag{12.2.8}$$

for all $t \in \mathbb{T}^+$. This formula is, of course, the density function of an exponential distribution with parameter β. A random variable T is said to have a Weibull distribution with parameter $\alpha > 0$ if its distribution function is

$$F(t) = 1 - \exp(-t^{\alpha}) \tag{12.2.9}$$

for $t \in \mathbb{T}^+$. It will be left as an exercise for the reader to show that the risk function for this distribution for $t > 0$ is

$$\theta(t) = \alpha t^{\alpha-1}. \tag{12.2.10}$$

Some of the ideas underlying models of competing risks may be illustrated in terms of various conceptual metaphors. Among the simplest of these metaphors is to consider some electronic device composed of many components, and suppose the design of the device is such that if one component fails, the device fails. Furthermore, suppose the device consists of $m \geq 2$ components and let the random variables $Y_1, Y_2, \ldots, Y_m$, which take values in the set $\mathbb{T}^+$, denote the life spans of the m components. It will be assumed that these random variables are independent with distribution functions $F_k(y)$ and survival functions $S_k(y) = 1 - F_k(y)$ for $k = 1, 2, \ldots, m$. These random variables and distributions are sometimes referred to as latent random variables and distributions.

Suppose the device is turned on at time $t = 0$ and observed until it fails. Let the random variable T denote the life span of the device. After a little reflection, it can be seen the random variable T is the minimum or smallest value of the random variables $Y_1, Y_2, \ldots, Y_m$. In symbols,

$$T = \min(Y_1, Y_2, \ldots, Y_m). \tag{12.2.11}$$

Thus, the failure of the device is due to the failure of component k if, and only if, $T = Y_k$. Let $S(t) = P[T > t]$ denote the survival function for the random variable T. Then, $T > t$, if, and only if, $Y_k > t$ for all $k = 1, 2, \ldots, m$. Therefore,

$$\begin{aligned} S(t) &= P[T > t] = P[Y_1 > t, Y_2 > t, \ldots, Y_m > t] \\ &= \prod_{k=1}^{m} P[Y_k > y] = \prod_{k=1}^{m} S_k(t), \end{aligned} \tag{12.2.12}$$

because, by assumption, the random variables $Y_1, Y_2, \ldots, Y_m$ are independent.

Before the formulas just derived can be useful in a computer implementation of the ideas underlying competing risks, it will be necessary to find some method or methods for parameterizing the latent distributions of a model for competing risks. A method that is often very useful in such endeavors is that of specifying a risk function of every latent distribution. Let $\theta_k(y)$ denote the risk function for latent distribution function $F_k(y)$ for $y \in \mathbb{T}^+$ and $k = 1, 2, \ldots, m$. Then from equations (12.2.5) and (12.2.12) it can be seen that

$$S(t) = \exp\left(-\int_0^t \theta(y)\, dy\right), \tag{12.2.13}$$

where $\theta(y) = \theta_1(y) + \theta_2(y) + \cdots + \theta_m(y)$. In particular, if there are positive constants β_k such that $\theta_k(y) = \beta_k$ for all $k = 1, 2, \ldots, m$, then this survival function reduces to that for the exponential distribution with parameter $\beta = \beta_1 + \beta_2 + \cdots + \beta_m$. As an exercise, a reader may also wish to consider a model of competing risks with latent risk functions of the Weibull type displayed in equation (12.2.10).

There is also another formula for a probability that arises, when considering models of competing risks, that will be very useful in subsequent sections of this chapter. In the model of competing risks considered above, let the function $G_k(t)$ denote the probability of the event $[T \leq t, T = Y_k]$. Observe that this event corresponds to the situation in which the time T of failure of the device is in the interval $(0, t]$ and the cause of failure is the failure of component k. In symbols,

$$G_k(t) = P[T \leq t, T = Y_k]. \tag{12.2.14}$$

When it is assumed that the latent distributions of the model are formulated in terms of risk functions, a formula for the probability $G_k(t)$ may be derived by an intuitive argument. Suppose at some time y in the interval

$(0,t]$ the system is still alive with probability $S(y)$ and, given this event, death or failure occurs sometime in the small time interval $(y, y+dy]$ with a conditional probability approximately equal to $\theta_k(y)\,dy$. By integrating over the interval $(0,t]$, it can be seen that

$$G_k(t) = \int_0^t S(y)\,\theta_k(y)\,dy. \tag{12.2.15}$$

In particular, if all the latent risks are constants β_k and $\beta = \beta_1 + \beta_2 + \cdots + \beta_m$, then this formula becomes

$$\begin{aligned} G_k(t) &= \int_0^t \exp(-\beta y)\,\beta_k dy \\ &= (1-\exp(-\beta t))\,\frac{\beta_k}{\beta} \end{aligned} \tag{12.2.16}$$

for $k = 1, 2, \ldots, k$ and $t \in \mathbb{T}^+$. As we shall see in what follows, formulas (12.2.15) and (12.2.16) are very useful when it is desired to parameterize transition functions for a class of stochastic processes known as semi-Markov processes.

An overview of Markov jump processes in continuous time and stationary laws of evolution was given in section 6.2, and in section 6.3 viewing these processes from the sample path perspective was outlined. Briefly, given some finite state space $\mathfrak{S} = (i_k \mid k = 1, 2, \ldots, r)$ for $r \geq 2$ and a matrix $\boldsymbol{Q} = (q_{ij})$ of rate parameters, it was indicated that if the process is in some state $\boldsymbol{X}_0 = i_0$ at time $t = 0$, then the waiting time to the next jump follows an exponential distribution with parameter $q_{i_0} = \sum_{j \neq i_0} q_{i_o j}$ Moreover, the conditional probability that the jump is to state $X_1 = j \neq i_0$ is $q_{i_o j}/q_{i_0}$. In general, sample path consisting of $n \geq 1$ jumps may be viewed as a set of ordered pairs $((i_k, t_k) \mid k = 0, 1, \ldots, n)$, where $X_k = i_k$ at jump k and $T_k = t_k$ for $k = 0, 1, \ldots, n$ is a realization of a random variable T_k denoting the waiting time to the next jump. For the class of Markov jump processes in continuous time, the distributions of the random variables T_k for $k = 0, 1, 2, \ldots, n$ are conditionally independent exponential random variables with scale parameters $q_k = \sum_{j \neq k} q_{kj}$ determined by the rate matrix $\boldsymbol{Q} = (q_{ij})$ and with conditional probabilities determined by $P[X_k = i_k \mid X_{k-1} = i_{k-1}] = q_{i_{k-1}, i_k}/q_{k-1}$ for $k = 1, 2, \ldots, n$.

When the evolution of a jump process in continuous time is formulated as a semi-Markov process, the distribution of the waiting in each state in the state space $\mathfrak{S}$ may be any distribution specified by the one who is formulating the process. A general way of looking at this class of processes is to image a $r \times r$ matrix of continuous and non-negative densities

$\boldsymbol{a}(t) = (a_{ij}(t))$ defined for all pairs of states $(i, j) \in \mathfrak{S} \times \mathfrak{S}$ and times $t \in \mathbb{T}^+$. Once this matrix is chosen, the construction of joint finite dimensional distributions of the process may proceed as follows. Let the symbol $\mathfrak{B}(n-1)$ denote any realization of the sample path random variables prior to the jump n. Then, the fundamental assumption underlying semi-Markov processes with stationary laws of evolution may be expressed as

$$\begin{aligned} &P[X_n = j, T_{n-1} \leq t \mid \mathfrak{B}(n-1)] \\ &= P[X_n = j, T_{n-1} \leq t \mid X_{n-1} = i] = \int_0^t a_{ij}(s)\, ds \qquad (12.2.17) \end{aligned}$$

for every pair of states $(i, j) \in \mathfrak{S} \times \mathfrak{S}$ and $t \in \mathbb{T}^+$. Observe that according to this equation, the future evolution of the process depends only on the state i visited in the previous jump $n-1$. Moreover, the waiting time to jump n depends only on state i and has the distribution function

$$\begin{aligned} F_i(t) &= \sum_{j \in \mathfrak{S}} P[X_n = j, T_{n-1} \leq t \mid X_{n-1} = i] \\ &= \sum_{j \in \mathfrak{S}} \int_0^t a_{ij}(s)\, ds. \qquad (12.2.18) \end{aligned}$$

From this equation, it is clear that the form of this distribution function depends entirely on row i of the matrix $\boldsymbol{a}(t) = (a_{ij}(t))$. In other words, the choices of the functions in this matrix determine the forms of these distribution functions for all states $i \in \mathfrak{S}$. A procedure, based on the ideas of competing risks, for construction this matrix will be discussed subsequently.

Before discussing this procedure, however, it is appropriate to discuss the finite dimensional distributions of the process based on the assumptions implicit in equation (12.2.17). Given that the process starts in state $i_0 \in \mathfrak{S}$, let $f_n(i_k, t_{k-1}, 1 \leq k \leq n \mid X_0 = i_0)$ denote the joint density of the collection of random variables $(X_k, T_{k-1} \mid k = 1, 2, \ldots, n)$ for states $i_0, i_1, \ldots, i_n$ in $\mathfrak{S}$ and times $t_0, t_1, \ldots, t_n$ in $\mathbb{T}^+$. Then, this joint density is given be the formula

$$f_n(i_k, t_{k-1}, 1 \leq k \leq n \mid X_0 = i_0) = \prod_{k=1}^{n} a_{i_{k-1}, i_k}(t_{k-1}). \qquad (12.2.19)$$

From this formula it is clear that probabilities are assigned directly to sample paths rather than at time points in the evolution of the process as was the case of Markov jump processes in continuous time, see equation (6.2.8).

As mentioned above, a very useful method of constructing the $r \times r$ matrix $\boldsymbol{a}(t) = (a_{ij}(t))$ of densities is to appeal to the theory of competing risks. Let $\boldsymbol{\Theta}(y) = (\theta_{ij}(y))$ denote a $r \times r$ matrix such that $\theta_{ij}(y) \geq 0$ for all pairs (i, j) and $y \in \mathbb{T}^+$. In order to preclude a transition to the same state, $\theta_{ii}(y) = 0$ for all $i \in \mathfrak{S}$ and $y \in \mathbb{T}^+$. Let

$$P[X_n = j, T_{n-1} \leq t \mid X_{n-1} = i] = A_{ij}(t) = \int_0^t a_{ij}(s)\, ds \qquad (12.2.20)$$

for each pair of states (i, j) and $y \in \mathbb{T}^+$. Then, given the matrix $\boldsymbol{\Theta}(y) = (\theta_{ij}(y))$ of latent risks, if $\theta_i(y)$ is defined as

$$\theta_i(y) = \sum_{j \in \mathfrak{S}} \theta_{ij}(y), \qquad (12.2.21)$$

then by applying the argument used in the derivation of formula (12.2.15), it follows that

$$A_{ij}(t) = \int_0^t S_i(y)\, \theta_{ij}(y)\, dy, \qquad (12.2.22)$$

where

$$S_i(t) = \exp\left(-\int_0^t \theta_i(y) dy\right). \qquad (12.2.23)$$

Therefore, the elements of the density matrix $\boldsymbol{a}(t)$ for this formulation are given by

$$a_{ij}(t) = \frac{dA_{ij}(t)}{dt} = S_i(t)\, \theta_{ij}(t) \qquad (12.2.24)$$

for pairs of states (i, j) and $y \in \mathbb{T}^+$.

If some state $i \in \mathfrak{S}$ is absorbing, then $\theta_{ij}(y) = 0$ for all $j \in \mathfrak{S}$ and $y \in \mathbb{T}^+$. If the state space $\mathfrak{S}$ contains $r_1 \geq 1$ absorbing states and $r_2 \geq 1$ transient states such that $r_1 + r_2 = r$, then the matrix of latent risks will have the partitioned form

$$\boldsymbol{\Theta}(y) = \begin{pmatrix} \mathbf{0}_{11} & \mathbf{0}_{12} \\ \boldsymbol{\Theta}_{21}(y) & \boldsymbol{\Theta}_{22}(y) \end{pmatrix} \qquad (12.2.25)$$

for all $y \in \mathbb{T}^+$, where $\mathbf{0}_{11}$ and $\mathbf{0}_{12}$ are $r_1 \times r_1$ and $r_1 \times r_2$ matrices of zeroes and the matrices $\boldsymbol{\Theta}_{21}$ and $\boldsymbol{\Theta}_{22}$ are matrices of latent risks of order $r_2 \times r_1$ and $r_2 \times r_2$, respectively.

In particular, if the matrix $\boldsymbol{\Theta}(y)$ is the constant matrix $\boldsymbol{\Theta} = (\theta_{ij})$, then formula (12.2.22) takes the form

$$A_{ij}(t) = (1 - \exp(-\theta_i t)) \frac{\theta_{ij}}{\theta_i}, \qquad (12.2.26)$$

for $\theta_i \neq 0$, and it follows that the distribution function of the sojourn time in state $i \in \mathfrak{S}$ is the exponential distribution function

$$F_i(t) = \sum_{j \in \mathfrak{S}} A_{ij}(t) = 1 - \exp(-\theta_i t). \tag{12.2.27}$$

As will be shown in subsequent sections of this chapter, formula (12.2.26) is very useful in formulating evolutionary models of populations in which ages of individuals are taken into account and it is assumed that risk functions are constant on time intervals of the form $(t, t+h] = [s \in \mathbb{T}^+ \mid t < s \leq t + h]$ for $h > 0$.

For those readers who may be interested in pursuing the literature in semi-Markov processes in greater depth, it is suggested that the books Barbu and Limnois (2008) and Mode and Sleeman (2000) be consulted for further references and applications of this class of stochastic processes.

12.3 Age Dependence and Types of Singles and Couples

In this section, the two sex population process introduced and studied in chapter 11 will be extended to include ages of individuals in an evolving population. The introduction of age into the formulation will in turn necessitate an extension of the definitions of the types of single individuals and couples that were defined in section 11.2. Although ages of individuals in a population evolve in continuous time, in any computer implementation of an evolutionary model that accommodates age, it will be necessary to partition ages into a finite number of disjoint sets which may be stored and easily manipulated in a computer. Let $r \geq 2$ denote a positive integer, which will be interpreted the maximum age considered in a computer implementation of an evolutionary process, and let $h > 0$ denote a unit of time under consideration such a day, month or year. Moreover, let the symbol $[x, x + h)$ denote the set $[t \in \mathbb{T}^+ \mid x \leq t < x + h]$, and let

$$([x, x + h) \mid x = 0, 1, 2, \ldots, r) \tag{12.3.1}$$

denote the set of intervals that partition ages into $r + 1$ disjoint sets. For example, the age class $[0, h)$ will denote infants born during some time period and the class $[r, r + h)$ denotes the oldest individuals in a simulated population. The age of individuals in the class $[x, x + h)$ will be denoted as x in the sense that their age at last birthday was x, where x is expressed in terms of the time unit under consideration. For the sake of brevity, the ages under consideration will be symbolized as $x = 0, 1, 2, \ldots, r$.

In populations of humans and their nearest relatives the great apes, there is a rather long period of adolescence before females and males, born during some time period, will be capable of reproducing. The age of sexual maturity for females and males in human populations is about 15 years. Consequently, when simulating the evolution of a human population with respect to two sexes, only the ages $x = 15, 16, \ldots, r$ of females and males will be eligible to participate in the formation of couples capable of reproducing. When other species are under consideration, however, the age of sexual maturity may be less or greater than that for humans. Consequently, in a computer implementation of the two sex evolutionary process under consideration, the symbol x_m will denote the age of sexual maturity so that the ages of individuals eligible for couple formation and reproduction will be denoted by $x = x_m, x_m + 1, \ldots, r$.

The inclusion of age and two sexes in an evolutionary model for some biological populations gives rise to the necessity of manipulating rather large arrays in a computer implementation of a process under consideration. Consequently, for the sake of concreteness and with a view towards limiting the size of computer arrays, genotypes will be considered only with respect to one autosomal locus with two alleles in some diploid population such as that for humans. Let the symbols A and a denote the alleles at some autosomal locus under consideration. Then, in what follows, the symbol $\mathfrak{T}$ with elements τ will denote the set (AA, Aa, aa) of three genotypes under consideration with respect to both the female and male populations. In particular, when the genotypes of either females or males are under consideration, the symbol τ for a genotypes will carry the subscript f or m to indicate whether an individual is female or male as denoted by the symbols τ_f and τ_m.

The symbol $\mathfrak{t}$ with or without subscripts will be used to designate the type of a single individual classified by genotype and age. A single female of genotype $\tau_f \in \mathfrak{T}$ of age $x = 0, 1, \ldots, r$ will be denoted by the symbol $\mathfrak{t}_f = (\tau_f, x)$. Similarly, a single male of genotype $\tau_m \in \mathfrak{T}$ and age $y = 0, 1, \ldots, r$ will be denoted by the symbol $\mathfrak{t}_m = (\tau_m, y)$. Let N_f denote the number of types of single females. Then, from the foregoing definitions, it can be seen that $N_f = 3 \times (r + 1)$. It can also be seen that the total number of types of single males is $N_m = 3 \times (r + 1)$. If, for example, the greatest age under consideration is $r = 100$ years, then the number of types of single females and males that need to be processed at each time period in a computer experiment simulating the evolution of a population would be $2 \times 3 \times (101) = 606$ when the time unit is one year. When planning a computer experiment, the processing of arrays of this size over a time period may give an investigator pause, particularly when a Monte Carlo simulation experiment is under consideration. However, if one is content to consider only a deterministic model embedded in an evolutionary stochastic process, then processing arrays of this size within acceptable periods of time is feasible with many present day desk top computers.

In section 11.2 couples were classified only with respect to the genotypes of the female and male in a couple. For example, if a female is of genotype τ_f and a male if is of genotype τ_m, then a couple with these genotypes was denoted by the symbol $\kappa = (\tau_f, \tau_m)$, which will also be useful when couples are classified with respect to the ages of the females and males. Let $\kappa = (\tau_f, \tau_m)$ denote a fixed couple type when the genotypes of the female and male are specified. Then a couple type in which the female is of age x and the male is of age y will be denoted by $\kappa_c = (\kappa, x, y)$. Sometimes couples of this type will also be denoted by the symbol $\kappa_c = (\tau_f, x; \tau_m, y)$. Let N_c be the number of types of couples. Then, in the formulation under consideration, if x_m is the age at which females and males reach sexual maturity, then this number is given by $N_c = 9 \times (r + 1 - x_m)^2$. When the fertile period of females in taken into account and $x_{\max}$ denotes the greatest age for which a female is capable of producing offspring, then the number of couple types that may produce offspring is given by the formula $N_c^* = 9 \times (x_{\max} + 1 - x_m) \times (r + 1 - x_m)$.

The number of couple types to be considered in a formulation can also be reduced if the condition that the ages of females and males in couples are highly correlated is taken into account. For example, let x denote the age of a female in a couple and let y denote the age of her potential mate. Then, by assuming that only those males of age y such that $| x - y | \leq 10$ would be acceptable as mates would significantly reduce the number of couple types to be considered in a formulation. In a subsequent section, this point will be discussed in greater depth.

These formulas are very useful when it desired to assess the number of couple types that one can feasibly be considered in a computer implementation of the formulation under consideration. For example, if $r = 100$, $x_{15} = 15$ and $x_{\max} = 50$, then it can be shown that $N_c = 66{,}564$ and $N_c^* = 27, 864$. On desk top computers with 3 to 6 giga bytes of memory, it may be feasible to entertain models with these parameter values if attention were restricted to a deterministic model embedded in a stochastic process. If one is patient and the computer has sufficient memory, then is may also be feasible to do Monte Carlo simulation experiments based on the stochastic version of the model. In human populations that existed 10,000 to 20,000 years ago, however, the maximum age in such population was probably about $r = 50$ years.

Therefore, for $r = 50$, these two numbers become $N_c = 11,664$ and $N_c^* = 11,664$. For many present day desk top computers, storing and manipulating arrays of this size would be feasible, particularly if attention were confined to a deterministic model embedded in a stochastic process. It is interesting to note that for these values of the parameters, the combined size of the arrays for types of single females and males would be $2 \times 3 \times 51 = 306$. For arrays of this size, it may also be feasible to do Monte Carlo simulation experiments based on the evolutionary stochastic process under consideration on existing desk top computers.

12.4 Altruism and Semi-Markovian Processes for Evolution of Single Individuals

In this section, semi-Markovian models for the evolution for single individuals in a population of females and males will be described for the age-dependent case. Among the principal tools that will be utilized in the construction of models will be that of competing risks of death, which, as will be shown, provide a useful framework for incorporating impact altruism and population density on mortality into a formulation. The set of transient states for the single population will be chosen as the set $\mathfrak{S}_{trans} = (x \mid x = 0, 1, 2, \ldots, r)$ of ages of individuals under consideration, which contains $r + 1$ states. To accommodate the death of individuals in the formulation, an absorbing state will be added to the state space of the process and will be denoted by d. Thus, in this formulation, the set $\mathfrak{S}_{absorb}$ of absorbing states is $\mathfrak{S}_{absorb} = (d)$.

In the age-dependent formulation under consideration, survivability will be a component of natural selection; therefore, latent risk functions will need to be specified for each sex, genotype $\tau \in \mathfrak{T}$ and age $x = 0, 1, \ldots, r-1$. Moreover, for any given age, the risk of death for females and males may differ. With this in mind, let $\theta\left(\tau_f, x\right)$ denote the latent risk function for a single female of genotype τ_f at age $x = 0, 1, 2, \ldots, r-1$. The risk function $\theta\left(\tau_m, x\right)$ for males of genotype τ_m at age x is defined similarly. It will be noted that these risk functions are not defined for age r, the greatest age considered in a computer implementation of the formulation. The reason for this omission is that, in the computer model, individuals of ages $x > r$ are not accounted for in the software. In the pursuit of realism however, in any actual computer implementation of the formulation, the number r will be chosen so large that ages greater than r can be neglected. The latent risk functions introduced in this paragraph are also known as intrinsic risks of death for each genotype and age so that in what follows the subscript *intrin* may be attached to a latent risk such as $\theta_{intrin}(\tau_f, x)$.

The population process under consideration will be self regulating or density dependent in the sense that a component of a risk of death for any individual also depends on total population size as was the case for the two sex process considered in chapter 11. Let the random function $Z_{tot}(t)$ denote total population size at some time t in a projection of the two sex and age dependent process under consideration. Just as was the case for ages of individuals, simulation of the evolution of a population in time will center around events occurring during time intervals of the form $[t, t+h) = [s \in \mathbb{T}^+ \mid t \leq s < t+h]$, where $h > 0$ is the time unit under consideration. Events occurring during this time interval will be discussed subsequently. Density dependence will be formulated in terms of a risk function corresponding to a Weibull survival function of the form $S(t) = \exp\left(-(\beta t)^{\alpha}\right)$, where α and β are positive parameters. It can be shown that the risk function corresponding to this survival function has the form

$$\theta(t) = \alpha\beta^{\alpha}t^{\alpha-1}. \tag{12.4.1}$$

Consider, for example, a single female of genotype τ_f and age x and let $\theta_{den}(\tau_f, x)$ denote the risk of death for this female due to population density. Then, under the assumption that risk of death of this female due to population density has form (12.4.1), it follows that $\theta_{den}(\tau_f, x)$ takes the form

$$\theta_{den}(t; \tau_f, x) = \alpha\beta^{\alpha}\left(Z_{tot}(t)\right)^{\alpha-1}, \tag{12.4.2}$$

during the time interval $[t, t+h)$, where the parameters α and β may depend on the genotype τ_f and age x. In actual computer implementation of this formula, to reduce the number of parameters that need to be considered, the α parameter will be assigned some value $\alpha \geq 2$ for all genotypes and ages. To further reduce the number of parameters that need to be specified in each computer experiment, the parameter β, which reflects the amount of resources available to the population, will be assumed to be constant over age intervals but may vary among genotypes to reflect the condition that some genotypes may be more competitive than others. For single males of genotype τ_m and age x, the risk function $\theta_{den}(t;\tau_m,x)$ will be defined as in (12.4.2) with possibly different β parameters to differentiate sexes.

For each female of genotype τ_f and age x, the total latent risk of death is

$$\theta_{tot}(t;\tau_f,x) = \theta_{intrin}(\tau_f,x) + \theta_{den}(t;\tau_f,x) \tag{12.4.3}$$

for ages $x = 0,1,2,\ldots,r-1$, and $\theta_{tot}(\tau_m,x)$, the total latent risk of death for males of genotype τ_m of age x is defined similarly. For reasons mentioned above, these risks are defined only for ages $x = 0,1,2,\ldots,r-1$ for both females and males. Let $\mathbf{\Theta}_d(t)$ denote a $r \times 1$ column vector with the elements

$$\mathbf{\Theta}_d(t;f) = (\theta_{tot}(t;\tau_f,x) \mid x = 0,1,2,\ldots,r-1). \tag{12.4.4}$$

Then, the $(r+1)\times(r+1)$ matrix $\mathbf{\Theta}(f,t)$ of latent risk for females for some time interval $[t,t+h)$ has the partitioned form

$$\mathbf{\Theta}(t;f) = \begin{pmatrix} 0 & \mathbf{0}_{1\times r} \\ \mathbf{\Theta}_d(t;f) & \mathbf{0}_{r\times r} \end{pmatrix}. \tag{12.4.5}$$

A corresponding matrix $\mathbf{\Theta}(t;m)$ of latent risk functions for males on this time interval is defined similarly. As only risks of death are under consideration, all risks for transitions among transient states have been set to 0. Of course however, those individuals who are of age x at time t but survive the time interval $[t,t+h)$ will be placed in the age class $x+1$ at time $t+h$.

In the formulation under consideration, it will be assumed that altruism, a concern of individuals for others in a population, will be expressed quantitatively in terms of reductions in intrinsic risks of death. One type of expression of altruism is the situation in which a member of a population observes that there is a predator in the area and warns others by vocal signals. Another type of expression of altruism is the attention, care and toleration that adults give to infants and adolescents, particularly among primates. At the outset, it is not entirely clear as to what methods may be used to quantify the expression of altruism. Suffice it to say that such methods are in the realm of statistical demography and are thus beyond the scope of this chapter. However, it is a matter of historical record that rates of mortality in man in some countries have declined, see, for example, the books Alderson (1981) and Keyfitz and Flieger (1968). It would also be of interest to consult the book Rogot, E. *et al.* (1988), which contains an analysis on mortality data for $1,000,000$ persons classified by demographic, social and economic factors. In the computer experiments reported in this chapter, none of the above cited data will actually be used but it will provide guidelines for calculating numerical versions of parameterized latent risk functions.

Given the background outlined above, the next step in developing algorithms to be used in Monte Carlo simulation experiments designed to study the evolution of population of single individuals is to set down procedures for calculating realizations of the sample functions of the process at times $t = 0, 1, 2, \ldots$ under consideration. At some time $t = 0, 1, 2, \ldots$ let the random function $X\left(t; \tau_f, x\right)$ denote the number of single females of genotype τ_f of age x in the population. The random function $Y\left(t; \tau_m, x\right)$ is defined similarly for single males. As in (12.4.3), let $\theta_{tot}\left(t; \tau_f, x\right)$ denote the total risk function for a single female of type $\mathfrak{t}_f = \left(\tau_f, x\right)$ at time t. Then, because the risk function $\theta_{tot}\left(t; \tau_f, x\right)$ is, by assumption constant on the interval $[t, t+h)$, it follows that the random time until death follows an exponential distribution with a parameter $\theta_{tot}\left(t; \tau_f, x\right)$. Consequently, the probability an individual of type $\mathfrak{t}_f = \left(\tau_f, x\right)$ is alive at time $t + h$ is $\exp(-\theta_{tot}\left(t; \tau_f, x\right) h)$ and $1-\exp(-\theta_{tot}\left(t; \tau_f, x\right) h)$ is the probability this type of individual is dead at time $t + h$.

Then, for ages $x = 1, 2, \ldots, r$, at time t let $X_S\left(t; \tau_f, x-1\right)$ denote the number of individuals of type $\mathfrak{t}_f = \left(\tau_f, x-1\right)$ who survive to age x during the time interval $[t, t+h)$. By assumption this random function is a realization from a conditional binomial distribution with index $X\left(t; \tau_f, x-1\right)$ and probability $p = \exp(-\theta_{tot}\left(t; \tau_f, x-1\right) h)$. Symbolically, this condition may be expressed as

$$X_S\left(t; \tau_f, x-1\right) \sim CBinom\left(X\left(t; \tau_f, x-1\right), p\right) \tag{12.4.6}$$

for $x = 1, 2, \ldots, r-1$. Realizations of the random function $Y_S\left(t; \tau_m, x-1\right)$ for males of type $\mathfrak{t}_m = \left(\tau_m, x-1\right)$ are also computed using a procedure similar to that in (12.4.6). Let the random function $X_D\left(t; \tau_f, x-1\right)$ denote the number of females of type $\mathfrak{t}_f = \left(\tau_f, x-1\right)$ who die during the time interval $[t, t+h)$. Then a realization of this random function may be computed using the formula

$$X_D\left(t; \tau_f, x-1\right) = X\left(t; \tau_f, x-1\right) - X_S\left(t; \tau_f, x-1\right), \tag{12.4.7}$$

and an analogous formula may be set down for males of type $\mathfrak{t}_m = \left(\tau_m, x-1\right)$.

Finally, the random function $X\left(t; \tau_f, 0\right)$, which denotes the total number of females of genotype τ_f born during the time interval $[t, t+h)$, would be computed by counting and summing the number of female births to all ages of mature females at the fertile ages $x = x_m, x_m + 1, \ldots, x_{\max}$. The number $Y\left(t; \tau_m, 0\right)$ of males of genotype τ_m born during the time interval $[t, t + h)$ would be computed using a similar procedure. In a subsequent section, the details underlying these computations will be given in more detail.

12.5 On an Age Dependent Couple Formation Process

Unlike the two sex process considered in chapter 11 section 11.3 in which age was not taken into account, in an age dependent population process the number of couples of each type formed during a time interval $[t, t + h)$ depends on the numbers of single females and males of ages greater than of equal to the age x_m of sexual maturity that survive this time interval. In this section, the couple formation process introduced in section 11.3 will be extended to the age dependent case. In what follows, these sets of females and males will be referred to as those eligible for the couple formation process. Although couple formation may take place throughout a time interval $[t, t+h)$, the task of writing of software to implement the age dependent process under consideration is greatly simplified if the formation of couples is taken into account only at the end of each time interval.

Let the random function $U_f\left(t; \tau_f, x\right)$ denote the frequency of type $\mathfrak{t}_f = \left(\tau_f, x\right)$ females in the eligible single female population who survive the time interval $[t, t + h)$. As in section 12.4, the random function $X_S\left(t; \tau_f, x\right)$ denotes the number of single females of this type who survive the time interval $[t, t + h)$ Then, given these definitions,

$$U_f\left(t; \tau_f, x\right) = \frac{X_S\left(t; \tau_f, x\right)}{X_S\left(t; \circ, \circ\right)} \tag{12.5.1}$$

for $X_S\left(t; \circ, \circ\right) > 0$ and $x \geq x_m$, where

$$X_S\left(t; \circ, \circ\right) = \sum_{\tau_f} \sum_{x=x_m}^{r} X_S\left(t; \tau_f, x\right). \tag{12.5.2}$$

If $X_S\left(t; \circ, \circ\right) = 0$, then $U_f\left(t; \tau_f, x\right) = 0$. The frequency $U_m\left(t; \tau_m, x\right)$ of eligible males of type $\mathfrak{t}_m = \left(\tau_m, x\right)$ for ages $x \geq x_m$ who survive the time interval $[t, t + h)$ is defined similarly.

The next step in the extending of the couple formation process described in section 11.3 is that of defining acceptance probabilities for females and males. Let $\alpha_f\left(\tau_f, x; \tau_m, y\right)$ denote the conditional probability that a female of type $\mathfrak{t}_f = \left(\tau_f; x\right)$ finds a male of type $\mathfrak{t}_m = \left(\tau_m, y\right)$ acceptable as a sexual partner. Similarly, let $\alpha_m(\tau_m, y; \tau_f, x)$ denote the conditional probability that a male of type $\mathfrak{t}_m = (\tau_m, y)$ finds a female of type $\mathfrak{t}_f = (\tau_f; x)$ acceptable as a sexual partner. Let $P_f(t; \tau_f.x)$ denote the probability that a single female of type $\mathfrak{t}_f = (\tau_f; x)$ has contact with some eligible single male during the time interval $[t, t+h)$. Then, by the law of total probability, this probability is given by

$$P_f(t; \tau_f, x) = \sum_{\tau_m} \sum_{y=x_m}^{r} U_m(t; \tau_m, y)\, \alpha_f(\tau_f; x; \tau_m, y)\,. \tag{12.5.3}$$

Similarly, let $P_m(t; \tau_f, y)$ denote the probability that a single male of type $\mathfrak{t}_m = (\tau_f, y)$ has contact with some eligible single female during the time interval $[t, t+h)$. Then,

$$P_m(t; \tau_f, y) = \sum_{\tau_f} \sum_{x=x_m}^{r} U_f(t; \tau_f, x)\, \alpha_m(\tau_m, y; \tau_f, x)\,. \tag{12.5.4}$$

Next let $\gamma_f(t; \tau_f, x; \tau_m, y)$ denote the conditional probability that an eligible female of type $\mathfrak{t}_f = (\tau_f, x)$ has contact with an eligible male of type $\mathfrak{t}_m = (\tau_m, y)$ during the time interval $[t, t+h)$. Then, by Bayes's theorem, it follows that

$$\gamma_f(t; \tau_f, x; \tau_m, y) = \frac{U_m(t; \tau_m, y)\, \alpha_f(\tau_f; x; \tau_m, y)}{P_f(t; \tau_f, x)}. \tag{12.5.5}$$

Similarly, let $\gamma_m(t; \tau_m, y; \tau_f, x)$ denote the conditional probability that an eligible male of type $\mathfrak{t}_m = (\tau_m, y)$ has contact with an eligible female of type $\mathfrak{t}_f = (\tau_f, x)$ during the time interval $[t, t+h)$. Then, as above

$$\gamma_m(t; \tau_m, y; \tau_f, x) = \frac{U_f(t; \tau_f, x)\, \alpha_m(\tau_m, y; \tau_f, x)}{P_m(t; \tau_f, y)}. \tag{12.5.6}$$

Just as in section 11.3, if the acceptance probabilities are constants, then these formulas reduce to

$$\gamma_f(t; \tau_f, x; \tau_m, y) = U_m(t; \tau_m, y) \tag{12.5.7}$$

and

$$\gamma_m(t; \tau_m, y; \tau_f, x) = U_f(t; \tau_f, x) \tag{12.5.8}$$

so that in this case the mating system is random.

The next step in the formulation of couple formation process for the age dependent case is that of parameterizing the acceptance probabilities in a form that will expedite the computer implementation of the process. With respect to a pair of genotypes (τ_f, τ_m), let $\alpha_f(\tau_f, \tau_m)$ be the conditional probability that a female of genotype τ_f finds a male of genotype τ_m

acceptable as a sexual partner, and let $\alpha_m(\tau_m;\tau_f)$ denote the conditional probability that a male of genotype τ_m finds a female of genotype τ_f acceptable as a sexual partner. Then, because age is a quantitative variable, a plausible and useful choice for the acceptance probability $\alpha_f(\tau_f;x;\tau_m,y)$ is

$$\alpha_f(\tau_f;x;\tau_m,y) = \alpha_f(\tau_f,\tau_m)\exp\left(-\beta_f \mid x-y \mid\right), \tag{12.5.9}$$

where β_f is a parameter such that $\beta_f \geq 0$. For those cases in which age in a significant factor in preferences for mates, the condition $\beta_f > 0$ would be satisfied. Similarly, the parameterization of the acceptance probability $\alpha_m(\tau_m,y;\tau_f,x)$ would take the form

$$\alpha_m(\tau_m,y;\tau_f,x) = \alpha_m(\tau_m,\tau_f)\exp\left(-\beta_m \mid x-y \mid\right), \tag{12.5.10}$$

where β_m is a parameter such that $\beta_m \geq 0$. To limit the number of couple types that need to be considered in a computer implementation of the process, one could specify in the software that the acceptance probabilities in (12.5.9) and (12.5.10) are 0 if $\mid x-y \mid > 10$. In human populations however, the vast majority of reproducing couples have ages such that $\mid x-y \mid \leq 10$, the use of this assumption would not seriously under estimate the number of births occurring during a time interval $[t,t+h)$. One could also consider cases in which (12.5.9) and (12.5.10) are 0 when $\mid x-y \mid > 5$, which would also greatly reduce the number of types of couples under consideration. When the age dependent model under consideration is applied to populations other than humans, however, the plausibility of assumptions of this type would need to be evaluated on a case by case basis.

In order to lighten the notation in what follows the set of ages for both females and males will be denoted by

$$\mathbb{A}_m = (x \mid x = x_m, x_m+1, \ldots, r), \tag{12.5.11}$$

and, as in the foregoing sections, the set of genotypes under consideration for both females and males will be denoted by $\mathfrak{T}$. For single females of type $\mathfrak{t}_f = (\tau_f, x)$, let

$$\gamma_f(t;\tau_f,x) = \left(\gamma_f(t;\tau_f,x;\tau_m,y) \mid \tau_m \in \mathfrak{T}, y \in \mathbb{A}_m\right) \tag{12.5.12}$$

denote a row vector of contact probabilities. And, similarly for single males of type $\mathfrak{t}_m = (\tau_m, y)$ let

$$\gamma_m(t;\tau_m,y) = \left(\gamma_m(t;\tau_m,y;\tau_f,x) \mid \tau_f \in \mathfrak{T}, x \in \mathbb{A}_m\right) \tag{12.5.13}$$

denote a row vector of contact probabilities for males.

For a single female of type $\mathfrak{t}_f = (\tau_f, x)$, let the random function

$$Z_f\left(t; \tau_f, x; \tau_m, y\right) \tag{12.5.14}$$

denote the number of single males of type $\mathfrak{t}_m = (\tau_m, y)$ selected as potential sexual partners during the time interval $[t, t+h)$, and let

$$\boldsymbol{Z}_f\left(t; \tau_f, x\right) = \left(Z_f\left(t; \tau_f, x; \tau_m, y\right) \mid \tau_m \in \mathfrak{T}, y \in \mathbb{A}_m\right) \tag{12.5.15}$$

denote a row vector of these random functions. Similarly, let $\boldsymbol{Z}_m\left(t; \tau_m, y\right)$ denote a similar vector of random functions for single males of type $\mathfrak{t}_m = (\tau_m, y)$ at time t. Given the number $X_S\left(t; \tau_f, x\right)$, it is assumed that the random vector $\boldsymbol{Z}_j\left(t; \tau_f, x\right)$ has a conditional multinomial distribution with index $X_S\left(t; \tau_f, x\right)$ and probability vector $\boldsymbol{\gamma}_f\left(t; \tau_f, x\right)$. In symbols,

$$\boldsymbol{Z}_f\left(t; \tau_f, x\right) \sim CMultinom\left(X_S\left(t; \tau_f, x\right), \boldsymbol{\gamma}_f\left(t; \tau_f, x\right)\right). \tag{12.5.16}$$

It will also be assumed that

$$\boldsymbol{Z}_m\left(t; \tau_m, y\right) \sim CMultinom\left(Y_S\left(t; \tau_m, y\right), \boldsymbol{\gamma}_m\left(t; \tau_m, y\right)\right). \tag{12.5.17}$$

Just as in section 11.3, the random function $N_C\left(t; \tau_f, x; \tau_m, y\right)$ will be defined at the potential number of couples of type $(\tau_f, x; \tau_m, y)$ formed during the time interval $[t, t+h)$. As this function cannot exceed the number of single females of type $\mathfrak{t}_f = (\tau_f, x)$ seeking single males of type $\mathfrak{t}_m = (\tau_m, y)$ and single males of type $\mathfrak{t}_m = (\tau_m, y)$ seeking single females of type $\mathfrak{t}_f = (\tau_f, x)$, it follows that

$$N_C\left(t; \tau_f, x; \tau_m, y\right) = \min\left(Z_f\left(t; \tau_f, x; \tau_m, y\right), Z_m\left(t; \tau_m, y; \tau_f, x\right)\right). \tag{12.5.18}$$

By following the ideas set forth in section 11.3, it can also be shown that

$$\sum_{\tau_m} \sum_{y=x_m}^{r} N_C\left(t; \tau_f, x; \tau_m, y\right) \leq X_S\left(t; \tau_f, x\right) \tag{12.5.19}$$

with probability one for all types $\mathfrak{t}_f = (\tau_f, x)$ of single females and time intervals of the form $[t, t+h)$. Similarly, it can be shown that

$$\sum_{\tau_f} \sum_{x=x_m}^{r} N_C\left(t; \tau_f, x; \tau_m, y\right) \leq Y_S\left(t; \tau_m, y\right) \tag{12.5.20}$$

with probability one for all types $\mathfrak{t}_m = (\tau_m, y)$ of single males and time intervals of the form $[t, t+h)$.

Let the random function $Z_C\left(t; \tau_f, x; \tau_m, y\right)$ denote the actual number of couples of type $\kappa_c = (t; \tau_f, x; \tau_m, y)$ formed during the time interval

$[t, t+h)$, and let $p(\kappa_c)$ denote a probability depending on the couple type κ_c. Then, given $N_C(t;\kappa_c)$, it will be assumed that

$$Z_C(t;\kappa_c) \sim CBinom(N_C(t;\kappa_c), p(\kappa_c)) \tag{12.5.21}$$

for all couple types κ_c. To reduce the number of parameters that need to be considered, in the computer experiments reported in subsequent sections of this chapter, it will be assumed that $p(\kappa_c) = p$, a constant, for all couple types κ_c.

For $\kappa_c = (t;\tau_f, x;\tau_m, y)$, let $X_C(t;\tau_f, x)$ denote the number of females of type $\mathfrak{t}_f = (\tau_f, x)$ who became members of a couple during the time interval $[t, t+h)$. Then, $X_C(t;\tau_f, x)$ is given by the sum

$$X_C(t;\tau_f, x) = \sum_{\tau_m \in \mathfrak{T}} \sum_{y=x_m}^{r} Z_C(t;\tau_f, x;\tau_m, y). \tag{12.5.22}$$

Similarly, let $Y_C(t;\tau_m, y)$ denote the number of males of type $\mathfrak{t}_m = (\tau_m, y)$ who became members of a couple during the time interval $[t, t+h)$. Then,

$$Y_C(t;\tau_m, y) = \sum_{\tau_f \in \mathfrak{T}} \sum_{x=x_m}^{r} Z_C(t;\tau_f, x;\tau_m, y). \tag{12.5.23}$$

As will be shown in a subsequent section, both these random functions will play a role in algorithms describing the evolution of the populations of single females and males.

12.6 A Semi-Markovian Model for Deaths, Dissolutions and Transitions Among Couple Types

A basic component of the two sex age dependent stochastic process under consideration is a module that accommodates death of either female or male of a couple as well as the possibility that a partnership may dissolve for reasons other than death. To accommodate these three possibilities, let $\mathfrak{S}_{c1}$ denote a set of three absorbing states

$$\mathfrak{S}_{c1} = (E_{1df}, E_{2dm}, E_{3dis}), \tag{12.6.1}$$

where E_{1df} and E_{3dm} denote, respectively, the death of the female or male member of a couple and E_{3dis} denotes the event that the partnership has been dissolved for other reasons. To simplify the notation in what follows, these three states will be denoted by E_j for $j = 1, 2, 3$. The set of transient states $\mathfrak{S}_{c2}$ of the process will be chosen as the set

$$\mathfrak{S}_{c2} = \mathfrak{K}_c = (\kappa_c = (\tau_f, x; \tau_m, y) \mid \tau_f \in \mathfrak{T}, \tau_m \in \mathfrak{T}, x \in \mathbb{A}, y \in \mathbb{A}) \qquad (12.6.2)$$

of couple types. As has been shown section 12.3, this set may contain a very large number of states so that in any practical computer implementation of the process it will be necessary to reduce the number of couple types under consideration.

The next step in the formulation of the model is that of introducing risk functions governing transitions from the set $\mathfrak{S}_{c2}$ of transient states to the set $\mathfrak{S}_{c1}$ of absorbing states. In this connection, all the risk function defined in section 12.4 for risks of death and population density will be in force in this section. To help fix ideas, suppose the number of couple types under consideration is $m \geq 2$ and let $\mathbf{\Theta}_{21}(t)$ be a $m \times 3$ matrix of latent risk functions governing transitions from the set $\mathfrak{S}_{c2}$ of transient states to the set $\mathfrak{S}_{c1}$ of absorbing states at time t. Then, during the time interval $[t, t+h)$, the row of this matrix corresponding to the couple type $\kappa_c = (\tau_f, x; \tau_m, y)$ would have the form

$$(\theta_{intrin}(\tau_f, x) + \theta_{den}(t; \tau_f, x), \theta_{intrin}(\tau_m, y) + \theta_{den}(t; \tau_m, y), \theta_{dis}), \qquad (12.6.3)$$

where, for the sake of simplicity, θ_{dis} is a constant risk for dissolution for all couple types $\kappa_c \in \mathfrak{S}_{c2}$. For the couple type $\kappa_c = (\tau_f, x; \tau_m, y)$, the total risk function at time t is

$$\begin{aligned}\theta_{tot}(t; \kappa_c) = \theta_{intrin}(\tau_f, x) + \theta_{den}(t; \tau_f, x) + \theta_{intrin}(\tau_m, y)\\ + \theta_{den}(t; \tau_m, y) + \theta_{dis}.\end{aligned} \qquad (12.6.4)$$

As all risk functions are assumed to be constant on the time interval $[t, t+h)$, it follows that the theory of competing risks with constant risk functions applies for transitions during this time interval. For $\kappa_c = (\tau_f, x; \tau_m, y)$ and $\kappa_c' = (\tau_f, x+1; \tau_m, y+1)$, let $\pi(t; \kappa_c, \kappa_c')$ denote the conditional probability of the transition $\kappa_c \to \kappa_c'$ during the time interval $[t, t+h)$. Then, because the risk functions are constant during this time interval, it follows that

$$\pi(t; \kappa_c, \kappa_c') = \exp(-\theta_{tot}(t; \kappa_c) h). \qquad (12.6.5)$$

For $j = 1, 2, 3$, let $\pi(t; \kappa_c, j)$ denote the probability of the transition $\kappa_c \to E_j \in \mathfrak{S}_{c1}$. Then, the conditional probability that the female member of the couple dies during the time interval $[t, t+h)$ is

$$\pi(t; \kappa_c, 1) = (1 - \exp(-\theta_{tot}(t; \kappa_c) h)) \frac{\theta_{intrin}(\tau_f, x) + \theta_{den}(t; \tau_f, x)}{\theta_{tot}(t; \kappa_c)}. \qquad (12.6.6)$$

A similar conditional probability $\pi\left(t;\kappa_c,2\right)$ that the male member of a couple dies during this time interval could also be written down. Finally, the conditional probability that the partnership dissolves for other reasons during this time interval is given by

$$\pi\left(t;\kappa_c,3\right)=\left(1-\exp\left(-\theta_{tot}\left(t;\kappa_c\right)h\right)\right)\frac{\theta_{dis}}{\theta_{tot}\left(t;\kappa_c\right)}. \tag{12.6.7}$$

At time t, let the random function $Z\left(t;\kappa_c\right)$ denote the number of couples of type κ_c in the population, and let the random function $W\left(t;\kappa_c,j\right)$ denote the number of transitions to absorbing state E_j, for $j=1,2,3$, during the time interval $[t,t+h)$. Then, let the random function $W\left(t;\kappa_c,\kappa_c'\right)$ denote the number of transitions of the form $\kappa_c\to\kappa_c'$ during this time interval, and let

$$\boldsymbol{W}\left(t;\kappa_c\right)=\left(W\left(t;\kappa_c,1\right),W\left(t;\kappa_c,2\right),W\left(t;\kappa_c,3\right),W\left(t;\kappa_c,\kappa_c'\right)\right) \tag{12.6.8}$$

denote a vector of these random functions. Next let

$$\boldsymbol{\Pi}\left(t;\kappa_c\right)=\left(\pi\left(t;\kappa_c,1\right),\pi\left(t;\kappa_c,2\right),\pi\left(t;\kappa_c,3\right),\pi\left(t;\kappa_c,\kappa_c'\right)\right) \tag{12.6.9}$$

denote a vector of conditional transition probabilities. Then, when computing Monte Carlo realizations of the process, it will be assumed that

$$\boldsymbol{W}\left(t;\kappa_c\right)\sim CMultinom\left(Z\left(t;\kappa_c\right),\boldsymbol{\Pi}\left(t;\kappa_c\right)\right). \tag{12.6.10}$$

When a couple of type $\kappa_c=\left(\tau_f,x;\tau_m,y\right)$ dissolves because of death of one of the partners, then the surviving partner returns to the single population. Similarly, a couple of this type dissolves for a reason other than death, then both partners return to the single population. It, therefore, becomes necessary to take such events into account when formulating procedures for the evolution of the population of single individuals. Let the random function $X_{DIS}\left(t;\kappa_c\right)$ denote the number of females who were members of a couple of type $\kappa_c=\left(\tau_f,x;\tau_m,y\right)$ at time t, but during the time interval $[t,t+h)$ this couple dissolves due to either the death of the male member of a couple or the dissolution of the couple for other reasons. Then, this random function is given by

$$X_{DIS}\left(t;\kappa_c\right)=W\left(t;\kappa_c,2\right)+W\left(t;\kappa_c,3\right). \tag{12.6.11}$$

Similarly, let the random function $Y_{DIS}\left(t;\kappa_c\right)$ denote the number of males who were members of a couple of type $\kappa_c=\left(\tau_f,x;\tau_m,y\right)$ at time t, but during the time interval $[t,t+h)$ this couple dissolves due to either the death of the female member of a couple or the dissolution of the couple for other reasons. By following a rationale similar to that used in (12.6.11), it follows that

$$Y_{DIS}\left(t;\kappa_c\right) = W\left(t;\kappa_c,1\right) + W\left(t;\kappa_c,3\right). \tag{12.6.12}$$

As will be shown in a subsequent section describing the stochastic evolution of the two sex age dependent process under consideration, these random functions will be included in the module describing the evolution of the single populations of females and males.

Let the random function $X_{DIS}\left(t;\tau_f,x\right)$ denote the number of females of type $\mathfrak{t}_f = (\tau_f, x)$ who were members of some couple type $\kappa_c = (\tau_f, x; \tau_m, y)$ at time t, but during the time interval $[t, t+h)$ couples of this type dissolved. Then,

$$X_{DIS}\left(t;\tau_f,x\right) = \sum_{\tau_m \in \mathfrak{T}} \sum_{y=x_m}^{r} X_{DIS}\left(t;\tau_f,x;\tau_m,y\right). \tag{12.6.13}$$

Similarly, let $Y_{DIS}\left(t;\tau_m,y\right)$ denote the number of males of type $\mathfrak{t}_m = (\tau_m, y)$, but during the time interval $[t, t+h)$ couples of this type dissolved. Then

$$Y_{DIS}\left(t;\tau_m,y\right) = \sum_{\tau_f \in \mathfrak{T}} \sum_{x=x_m}^{r} Y_{DIS}\left(t;\tau_f,x;\tau_m,y\right). \tag{12.6.14}$$

As will be shown in a subsequent section, these random functions will play a role in describing the evolutionary dynamics of an age dependent population with two sexes.

12.7 Gamete, Genotypic and Offspring Distributions for Each Couple Type

Just as in section 11.5, the case of an autosomal locus with two alleles A and a will be under consideration in this section so that most of the results described in section 11.5 for the set $\mathfrak{T} = (AA, Aa, aa)$ of three genotypes remain in force in this section. In this section however, it will be convenient to denote a couple of type $\kappa_c = (\tau_f, x; \tau_m, y)$ in the alternative form $\kappa_c = (\tau_f, \tau_m; x, y)$ to indicate that the focus of attention is on the genotypes of female and male in a couple. In what follows, it will be assumed that the gamete distribution of any individual of genotype $\tau \in \mathfrak{T}$ does not depend on the age of this individual. Similarly, for a couple of type $\kappa_c = (\tau_f, \tau_m; x, y)$, it will be assumed that genotypic distribution depends only the genotypes

(τ_f, τ_m) of the female and male members of the couple and not on their ages (x, y).

For any genotype $\tau \in \mathfrak{T}$, the set of possible gametes will be denoted by $\mathfrak{G}_g = (A, a)$, and let $p_g(\tau; \nu)$ denote the probability that an individual of genotype τ produces a gamete of type $\nu \in \mathfrak{G}_g$. Like the case described in section 11.5, the gamete distribution for each genotype $\tau \in \mathfrak{T}$ depends on the matrix of mutation probabilities

$$\mathfrak{M} = \begin{pmatrix} \mu_{11} & \mu_{12} \\ \mu_{21} & \mu_{22} \end{pmatrix}, \tag{12.7.1}$$

and for each genotype $\tau \in \mathfrak{T}$, the probability $p_g(\tau; \nu)$ is a function of elements of the matrix $\mathfrak{M}$. All these probabilities are listed in Table 11.5.1, and from this table it can be seen that if $\tau = AA$ and $\nu = A$, then $p_g(\tau; \nu) = \mu_{11}$ and if $\nu = a$, then $p_g(\tau; \nu) = \mu_{12}$. By continuing in this way and using Table 11.5.1, formulas for the collection of six probabilities $\left(p_g(\tau; \nu) \mid \tau \in \mathfrak{T}, \nu \in \mathfrak{G}_g\right)$ could be set down as functions of the elements of the mutation matrix in (12.7.1).

For a couple of type $\kappa_c = (\tau_f, \tau_m; x, y)$, let $p(\tau_f, \tau_m; \tau)$ denote the conditional probability that they produce an offspring of genotype $\tau = (\nu, \nu')$, where ν and ν' are, respectively, the alleles contributed by the female and male parents. For each fixed couple type κ_c, the collection of probabilities $(p(\tau_f, \tau_m; \tau) \mid \tau \in \mathfrak{T})$ will be called the genotypic distribution for this couple type. Under the assumption that the gametes of the female and male parents combine independently in the offspring, it follows that

$$p(\tau_f, \tau_m; \tau) = p_g(\tau_f; \nu)\, p_g(\tau_m; \nu'). \tag{12.7.2}$$

Rather than going through the tedious procedure of deriving an explicit algebraic form for each of the probabilities in (12.7.2), in any computer implementation of the model, it is more useful to write software to compute these probabilities for all combinations of $(\tau_f, \tau_m; \tau)$. Formally, this collection of genotypic distributions may be thought of as a 9×3 array

$$(p(\tau_f, \tau_m; \tau) \mid \tau_f \in \mathfrak{T}, \tau_m \in \mathfrak{T}, \tau \in \mathfrak{T}). \tag{12.7.3}$$

For a couple of type $\kappa_c = (\tau_f, \tau_m; x, y)$, let the random variable $N(\kappa, x)$, taking values in the set of non-negative integers $(n \mid n = 0, 1, 2, \ldots)$, denote the number of offspring produced by this type couple during the time interval $[t, t+h)$. It will be assumed that this random variable has a Poisson distribution with parameter $\lambda(\kappa, x)$, depending on the age of the female x and the genotypes $\kappa = (\tau_f, \tau_m)$ of the parents. The parameter $\lambda(\kappa, x)$

is a component of natural selection in the sense that the reproductive success of each couple type depends on the genotypes of the parents. It also reflects a component of reproductive physiology of human female in that the expected number of offspring that she produces at age x decreases as a function of x. A helpful way of interpreting this function is to note that

$$\lambda(\kappa) = \sum_{x=x_m}^{x_{\max}} \lambda(\kappa, x) \tag{12.7.4}$$

is the expected number of offspring produced throughout the fertile period of the female, given that she marries at age x_m and survives along with her husband to age $x_{\max}$, the end of the fertile period of a female. In demography, this function is sometimes called the total fertility rate.

In the presence of mutation, a couple with genotypes $\kappa = (\tau_f, \tau_m)$ may produce an offspring with any of the three genotypes under consideration. Accordingly, let the vector $\boldsymbol{\xi} = (\xi_1, \xi_2, \xi_3)$ denote a generalized Bernoulli indicator function which indicates the genotype of the offspring. For example, if $\boldsymbol{\xi} = \boldsymbol{\varepsilon}_1 = (1, 0, 0)$, then an offspring is of genotype AA. In general $\boldsymbol{\varepsilon}_k$ denotes a vector indicator such there is a 1 in position k and zeroes elsewhere. To lighten the notation, let the symbols $\tau_1 = AA, \tau_2 = Aa$ and $\tau_3 = aa$ denote the three genotypes under consideration, and let $\boldsymbol{\tau} = (\tau_1, \tau_2, \tau_3)$ denote a vector of the three genotypic symbols. Next, let the vector valued random function $\boldsymbol{V}(t; \tau_f, \tau_m; x, y; \boldsymbol{\tau})$ denote the numbers of offspring of each of the genotypes in the vector $\boldsymbol{\tau}$ produced by a couple of type $\kappa_c = (\tau_f, \tau_m; x, y)$ during the time interval $[t, t+h)$. Then, this vector is the random sum

$$\boldsymbol{V}(t; \tau_f, \tau_m; x, y; \boldsymbol{\tau}) = \sum_{\nu=1}^{N(\kappa,x)} \boldsymbol{\xi}_\nu, \tag{12.7.5}$$

where $\boldsymbol{\xi}_\nu$ for $\nu = 1, 2, \ldots, N(\kappa, x)$ is a collection of generalized conditionally independent Bernoulli indicator vectors. Of course, if $N(\kappa, x) = 0$, then $\boldsymbol{V}(t; \tau_f, \tau_m; x, y; \boldsymbol{\tau}) = \mathbf{0}$, the zero vector. Another way of viewing the random vector in (12.7.5) is that it has a conditional multinomial distribution with index $N(\kappa, x)$ and probability vector

$$\boldsymbol{p}(\tau_f, \tau_m; \boldsymbol{\tau}) = (p(\tau_f, \tau_m; \tau_1), p(\tau_f, \tau_m; \tau_2), p(\tau_f, \tau_m; \tau_3)), \tag{12.7.6}$$

see (12.7.3).

When one is considering an evolutionary stochastic process at the population level, it will be necessary to consider the number of couples of type $\kappa_c = (\tau_f, \tau_m; x, y)$ present in a population at any time t. Let the random function $Z(t; \kappa_c)$ denote the number of couples of type κ_c present in a population at time t, and let the random vector $\boldsymbol{V}_C\left(t;\kappa_c;\boldsymbol{\tau}\right)$ denote the numbers of offspring of each of the three genotypes under consideration produced by the $Z\left(t;\kappa_c\right)$ couples during the time interval $[t,t+h)$. Then, $\boldsymbol{V}_C\left(t;\kappa_c;\boldsymbol{\tau}\right)$ is the random sum

$$\boldsymbol{V}_C\left(t;\kappa_c;\boldsymbol{\tau}\right)=\sum_{\nu=1}^{Z(t;\kappa_c)}\boldsymbol{V}_\nu\left(t;\tau_f,\tau_m;x,y;\boldsymbol{\tau}\right),\tag{12.7.7}$$

where $\boldsymbol{V}_\nu\left(t;\tau_f,\tau_m;x,y;\boldsymbol{\tau}\right)$ for $\nu=1,2,\ldots,Z\left(t;\kappa_c\right)$ are conditionally independent vector random variables with the same distribution as the random vector in (12.7.5). If $Z\left(t;\kappa_c\right)=0$, then $\boldsymbol{V}_C\left(t;\kappa_c;\boldsymbol{\tau}\right)=\mathbf{0}$.

As was indicated in section 12.4, to complete the description of the computational procedures used in the studying the evolution of the singles population, it will be necessary to compute realizations of the random functions $X_B\left(t;\tau_f,0\right)$ and $Y_B\left(t;\tau_m,0\right)$ denoting, respectively, the number of females and males born during the time interval $[t,t+h)$ with genotypes τ_f and τ_m. From the random vector in (12.7.7), it can be seen that the random vector

$$\boldsymbol{V}_C\left(t;\tau_f,\tau_m;x,\circ;\boldsymbol{\tau}\right)=\sum_{y=x_m}^{x_{\max}}\boldsymbol{V}_C\left(t;\tau_f,\tau_m;x,y;\boldsymbol{\tau}\right)\tag{12.7.8}$$

is the total number of offspring of each of the three genotypes in the vector $\boldsymbol{\tau}=\left(\tau_1,\tau_2,\tau_3\right)$ born during the time interval $[t,t+h)$ to parents with genotypes $\kappa=\left(\tau_f,\tau_m\right)$ and mothers of age $x=x_m,x_m+1,\ldots,x_{\max}$. Therefore, the total number of offspring born during this time interval with the three genotypes in the vector $\boldsymbol{\tau}=\left(\tau_1,\tau_2,\tau_3\right)$ is given by the vector valued random function

$$\boldsymbol{V}_C\left(t;\tau_f,\tau_m;\circ,\circ;\boldsymbol{\tau}\right)=\sum_{x=x_m}^{x_{\max}}\boldsymbol{V}_C\left(t;\tau_f,\tau_m;x,\circ;\boldsymbol{\tau}\right).\tag{12.7.9}$$

This random vector contains the total number of offspring of each of the three genotypes born during the time interval $[t,t+h)$ to couples with genotypes $\left(\tau_f,\tau_m\right)$ but, to accommodate the presence of two sexes in the formulation, each of the three numbers needs to be partitioned into females and males. Let p_f denote the probability an offspring is female and let $p_m=1-p_f$ denote the probability an offspring is male. Next rewrite the random vector $\boldsymbol{V}_C\left(t;\tau_f,\tau_m;\circ,\circ;\boldsymbol{\tau}\right)$ in its component form:

$$\boldsymbol{V}_C\left(t;\tau_f,\tau_m;\circ,\circ;\boldsymbol{\tau}\right)=\left(V_C\left(t;\tau_f,\tau_m;\circ,\circ;\tau_\nu\right)\mid\nu=1,2,3\right).\tag{12.7.10}$$

Let the random function $X\left(t;\tau_f,\tau_m;\tau_\nu,0\right)$ denote the number of females of genotype τ_ν born to a couple with genotypes $\left(\tau_f,\tau_m\right)$ during the time interval $[t, t+h)$. Then, for $\nu = 1, 2, 3$, given $V_C\left(t; \tau_f, \tau_m; \circ, \circ; \tau_\nu\right)$, the conditional distribution of $X\left(t; \tau_f, \tau_m; \tau_\nu, 0\right)$ is

$$X_B\left(t; \tau_f, \tau_m; \tau_\nu, 0\right) \sim CBinom\left(V_C\left(t; \tau_f, \tau_m; \circ, \circ; \tau_\nu\right), p_f\right), \qquad (12.7.11)$$

and, moreover the number of males of type $\mathfrak{t}_m = \left(\tau_\nu, 0\right)$ born to couples of this type$\left(\tau_f, \tau_m\right)$ during the this time interval is

$$Y_B\left(t; \tau_f, \tau_m; \tau_\nu, 0\right) = V_C\left(t; \tau_f, \tau_m; \circ, \circ; \tau_\nu\right) - X_B\left(t; \tau_f, \tau_m; \tau_\nu, 0\right). \qquad (12.7.12)$$

Therefore, the total number of females of type $\mathfrak{t}_f = \left(\tau_f, 0\right)$ born to all couples during the time interval $[t, t+h)$ is

$$X_B\left(t; \tau_f, 0\right) = \sum_{\tau'_f} \sum_{\tau'_m} X_B\left(t; \tau'_f, \tau'_m; \tau_f, 0\right) \qquad (12.7.13)$$

Similarly,

$$Y_B\left(t; \tau_m, 0\right) = \sum_{\tau'_f} \sum_{\tau'_m} Y_B\left(t; \tau'_f, \tau'_m; \tau_m, 0\right) \qquad (12.7.14)$$

is the total number of males of $\mathfrak{t}_m = \left(\tau_m, 0\right)$ born to all couples during the this time interval.

Equation (12.7.7) is mathematically correct, but from the computational point of view, it entails many calls for realizations of Poisson random variables with parameters $\lambda\left(\kappa, x\right)$, which depend only of the genotypes $\kappa = \left(\tau_f, \tau_m\right)$ of the female and male members of a couple as well as on the age x of the female. Consequently, for couples of type $\kappa_c = \left(\tau_f, \tau_m; x, y\right)$, if one sums over the ages y of the males in a couple type, then the number of calls for realizations of Poisson random variable could be reduced. For example, consider the random function $Z\left(t; \kappa_c\right)$, denoting the number of couples of type $\kappa_c = \left(\tau_f, \tau_m; x, y\right)$ in the population at time t, and define the random function $Z\left(t; \left(\tau_f, \tau_m; x, \circ\right)\right)$ by the sum

$$Z\left(t; \tau_f, \tau_m; x, \circ\right) = \sum_{y=x_m}^{r} Z\left(t; \left(\tau_f, \tau_m; x, y\right)\right). \qquad (12.7.15)$$

Given $Z\left(t; \tau_f, \tau_m; x, \circ\right)$, let $N_v\left(\kappa, x\right)$ for $\nu = 1, 2, \ldots, Z\left(t; \left(\tau_f, \tau_m; x, \circ\right)\right)$ denote a collection of conditionally independent Poisson random variables with a common parameter $\lambda\left(\kappa, x\right)$, where $\kappa = \left(\tau_f, \tau_m\right)$. Then let

$$H\left(t; \kappa, x\right) = \sum_{\nu=1}^{Z\left(t; \tau_f, \tau_m; x, \circ\right)} N_v\left(\kappa, x\right) \qquad (12.7.16)$$

Given $H\left(t;\kappa,x\right)$, the vector valued random function in (12.7.8) may be computed as a realization from a multinomial distribution with index $H\left(t;\kappa,x\right)$ and probability vector $\boldsymbol{p}\left(\tau_f,\tau_m;\boldsymbol{\tau}\right)$, see (12.7.6). In symbolic form

$$\boldsymbol{V}_C\left(t;\tau_f,\tau_m;x,\circ;\boldsymbol{\tau}\right)\sim CMultinom(H\left(t;\kappa,x\right),\boldsymbol{p}\left(\tau_f,\tau_m;\boldsymbol{\tau}\right)) \quad (12.7.17)$$

It can be shown that the random variable $H\left(t;\kappa,x\right)$ in (12.7.16) has a Poisson distribution with parameter $Z\left(t;\tau_f,\tau_m;x,\circ\right)\lambda\left(\kappa,x\right)$. Thus, a realization of this random variable may be computed as a sample of one from a Poisson distribution with this parameter, but, if $Z\left(t;\tau_f,\tau_m;x,\circ\right)$ is large, then a realization of $H\left(t;\kappa,x\right)$ may be computed by using a central limit theorem approximation. From these remarks it is clear that, by using (12.7.16) and (12.7.17), many calls for realizations of Poisson random variables may be avoided.

The feasibility of implementing the ideas outlined in this section in computer software also depends on the number of couple types under consideration. It is, therefore, of interest to reduce the number of couple types under consideration, using some plausible scheme. For any couple type $\kappa_c=\left(\tau_f,\tau_m;x,y\right)$ and for fixed values of in the triple $\left(\tau_f,\tau_m;x\right)$, consider the number of ages y of males such that $\mid x-y\mid\leq 5$. Then, for a given x, the values of y that satisfy this condition would have the form $y=x+\nu$ or $y=x-\nu$ for $\nu=0,1,2,\ldots,5$. Therefore, for a fixed x, the number of y values that satisfy the condition is 11. Observe that if $x=x_m$ or x is near x_m, fewer than 11 values of $y\geq x_m$ would satisfy this condition, but, the finding of these smaller numbers will be left as an exercise for the reader. In all cases, the condition $\mid x-y\mid\leq 5$, would significantly reduce the number of summands needed to compute a realization of the random function in (12.7.13) for each triplet $\left(\tau_f,\tau_m;x\right)$.

If, for example, $r=100, x_m=15$ and $x_{\max}=50$, then for $x=x_m,x_m+1,\ldots,x_{\max}$, the number of couple types $\kappa_c=\left(\tau_f,\tau_m;x,y\right)$ satisfying the condition $\mid x-y\mid\leq 5$ would be about $9\times 36\times 11=3,564$. If those couple types such that $x>x_{\max}$ were ignored, because, by assumption, such couples would produce no offspring, then this number of couple types would be manageable on many existing desk top computer, particularly if attention were confined to the embedded deterministic model. Moreover, if a desk top computer has memory of 3 to 6 giga bytes, then it would feasible to entertain this many couple types on such a desk top computer in stochastic simulation experiments. On the other hand, if a cluster of computers connected in a network were available, then entertaining a model with $3,564$ couple types in stochastic simulation would be even more feasible, even though the time taken to complete such an experiment may be quite large.

12.8 Overview of Stochastic Population Process with Two Sexes and Age Dependence

The purpose of this section is to set down a set of equations, governing the evolutionary dynamics of the two sex age dependent process under consideration. These equations will be partitioned into three sets, consisting of single females and males of each type as well as the number of couples of each type present in the population at any time $t \in (kh \mid k = 0, 1, 2, \ldots)$, the set of time points under consideration. This set of equations will not only form the basis for constructing software to compute Monte Carlo realizations of the process but will also provide a framework for the derivation of a system of non-linear difference equations embedded in the stochastic process. Even though this system of equations may be represented in a succinct symbolic form, any computer implementation of the process would involve many equations, particularly the set of equations describing the evolution of couple types.

For example, let the random function $X(t + h; \tau_f, 0)$ denote the number of female infants of type $\mathfrak{t}_f = (\tau_f, 0)$born during the time interval $[t, t + h)$ who survive to time $t + h$. Then, according to the results presented in section 12.7, it follows that

$$X(t + h; \tau_f, 0) = X_B(t; \tau_f, 0), \tag{12.8.1}$$

for every genotype τ_f, see (12.7.13). Observe that in the formulation under consideration, deaths to infants born during the time interval $[t, t + h)$ will not be taken into account until the next time interval $[t+h, t+2h)$. For ages $x = 1, 2, \ldots, r$, let the random function $X(t + h; \tau_f, x)$ denote the number of single females of type $\mathfrak{t}_f = (\tau_f, x)$ who are alive at time $t + h$. Then, for ages $x = 1, 2, \ldots, x_m - 1$ for which single individuals are not eligible to become members of a couple, it follows that

$$X(t + h; \tau_f, x) = X_S(t; \tau_f, x - 1), \tag{12.8.2}$$

see (12.4.6). For ages $x = x_m, x_m + 1, \ldots, r$ however, this random function consist of those individuals of type τ_f who were alive at time t and survived to time $t + h$ minus those who became members of couples during the time interval $[t, t + h)$. To this difference the number of those who returned to single population during the time interval $[t, t + h)$ because of couple dissolution must be added. Thus, for ages $x = x_m, x_m + 1, \ldots, r$ and all genotypes τ_f

$$X(t+h;\tau_f,x) = X_S(t;\tau_f,x-1) - X_C(t;\tau_f,x-1) + X_{DIS}(t;\tau_f,x-1), \tag{12.8.3}$$

see (12.5.22) and (12.6.13). To avoid ambiguities, let $X_C(t;\tau_f,x-1) = 0$ and $X_{DIS}(t;\tau_f,x-1) = 0$ if $x = x_m$.

For the sake of completeness and to aid in the writing of software, the analogous equations for single males will also be presented. Thus for single males of type $\mathfrak{t}_m = (\tau_m, 0)$, it follows that

$$Y(t+h;\tau_m,0) = Y_B(t;\tau_m,0), \tag{12.8.4}$$

see (12.7.14). Similarly, for ages $x = 1, 2, \ldots, x_m - 1$

$$Y(t+h;\tau_m,x) = Y_S(t;\tau_m,x-1) \tag{12.8.5}$$

for single males of type $\mathfrak{t}_m = (\tau_m, x)$. But, for ages $x = x_m, x_m + 1, \ldots, r$

$$Y(t+h;\tau_m,x) = Y_S(t;\tau_m,x-1) - Y_C(t;\tau_m,x-1) + Y_{DIS}(t;\tau_m,x-1) \tag{12.8.6}$$

for single males of type $\mathfrak{t}_m = (\tau_m, x)$.

Let the random function $Z(t+h;\kappa_c)$ denote the number of couples of type $\kappa_c = (\tau_f, \tau_m; x, y)$ present in the population at time $t+h$. Then let $\kappa_c' = (\tau_f, \tau_m; x-1, y-1)$. Then, for all couples of type κ_c, it follows that

$$Z(t+h;\kappa_c) = Z_C(t;\kappa_c') + W(t;\kappa_c',\kappa_c), \tag{12.8.7}$$

where $Z_C(t;\kappa_c')$ is the number of couples of type κ_c' formed during the time interval $[t, t+h)$ from females and males of types $\mathfrak{t}_f = (\tau_f, x-1)$ and $\mathfrak{t}_m = (\tau_m, y-1)$ at time t and $W(t;\kappa_c',\kappa_c)$ is the number of couples of type κ_c' at time t who made the transition $\kappa_c' \rightarrow \kappa_c$ during the time interval. For the definition of the random function $Z_C(t;\kappa_c')$ see (12.5.22) and consult (12.6.8) for the definition of the random function $W(t;\kappa_c',\kappa_c)$. Equation (12.8.7) is valid for all ages such that $x \geq x_m + 1$ and $y \geq x_m + 1$ and all pairs of genotypes (τ_f, τ_m).

In section 12.7, it was noted that, in order to minimize the number of couple types, the set of couple types considered in a computer implementation of the process should contain only females of ages $x = x_m, \ldots, x_{\max}$ and ages of males such that $\mid x - y \mid \leq 5$. If this were the case, then the above equations would have to be modified. A simple way of modifying these equations would be that assigning all couple types such that the ages x and y of females and males exceed the age $x_{\max}$ to the respective single classes of females and males but the formal details entailed in such assignments will not be discussed further here.

12.9 Overview of Non-Linear Difference Equations Embedded in the Stochastic Population Process

In this section, the non-linear difference equations embedded in the stochastic population process will be derived from the stochastic evolutionary equations described in section 12.8. Just as in previous chapters, these deterministic equations will be derived by taking the conditional expectation of some random function of process at time $t+h$, given the evolution of the process up to time t. Let the symbol $\Xi(t)$ stand for the evolution of the process up to time t. To begin the discussion of the procedure for deriving the deterministic non-linear difference equations embedded in the stochastic process with reference to equation (12.8.1) consider the conditional expectation

$$E\left[X\left(t+h;\tau_f,0\right) \mid \Xi\left(t\right)\right]. \tag{12.9.1}$$

As this random function is a non-linear function of the sample functions of the process at time t, it will be estimated from estimates of the sample functions at time t. From (12.7.11) it can be seen that given $V_C\left(t;\tau_f,\tau_m;\circ,\circ;\tau_\nu\right)$, the conditional expectation of $X_B\left(t;\tau_f,\tau_m;\tau_\nu,0\right)$ is

$$E\left[X_B\left(t;\tau_f,\tau_m;\tau_\nu,0\right) \mid V_C\left(t;\tau_f,\tau_m;\circ,\circ;\tau_\nu\right)\right] = V_C\left(t;\tau_f,\tau_m;\circ,\circ;\tau_\nu\right)p_f. \tag{12.9.2}$$

Let $\Xi_Z\left(t\right)$ denote the collection

$$\Xi_Z\left(t\right) = \left(Z\left(t;\tau_f,\tau_m;x,y\right) \mid x \in A_m, y \in A_m\right), \tag{12.9.3}$$

where A_m is the set $(x \mid x = x_m,\ldots,r)$. To simplify the notation let

$$Z\left(t;\tau_f,\tau_m;x,\circ\right) = \sum_{y=x_m}^{r} Z\left(t;\tau_f,\tau_m;x,y\right), \tag{12.9.4}$$

and let

$$\Psi\left(t;\tau_f,\tau_m;\tau_\nu\right) = \sum_{x=x_m}^{x_{\max}} Z\left(t;\tau_f,\tau_m;x,\circ\right)\lambda\left(\tau_f,\tau_m;x\right)p\left(\tau_f,\tau_m;\tau_\nu\right). \tag{12.9.5}$$

Then,

$$E\left[V_C\left(t;\tau_f,\tau_m;\circ,\circ;\tau_\nu\right) \mid \Xi_Z\left(t\right)\right] = \Psi\left(t;\tau_f,\tau_m;\tau_\nu\right). \tag{12.9.6}$$

Therefore,

$$E[X_B\left(t;\tau_f,\tau_m;\tau_\nu,0\right) \mid \Xi_Z\left(t\right)] = \Psi\left(t;\tau_f,\tau_m;\tau_\nu\right)p_f. \tag{12.9.7}$$

Consequently,

$$E\left[X_B\left(t;\tau_f,0\right)\mid\Xi_Z\left(t\right)\right]=\Upsilon\left(t;\tau_f\right)p_f, \tag{12.9.8}$$

where

$$\Upsilon\left(t;\tau_\nu\right)=\sum_{\tau_f'}\sum_{\tau_m'}\Psi\left(t;\tau_f',\tau_m';\tau_\nu\right). \tag{12.9.9}$$

Similarly,

$$E\left[Y_B\left(t;\tau_m,0\right)\mid\Xi_Z\left(t\right)\right]=\Upsilon\left(t;\tau_m\right)p_m. \tag{12.9.10}$$

Let

$$\widehat{\Xi}\left(t\right)=\left(\widehat{Z}\left(t;\tau_f,\tau_m;x,y\right)\mid x\in A_m,y\in A_m\right) \tag{12.9.11}$$

denote a collection of estimates of the random functions in the collection $\Xi\left(t\right)$ at time t and let $\widehat{\Upsilon}\left(t;\tau_\nu\right)$ denote an estimate of the random function $\Upsilon\left(t;\tau_\nu\right)$ based on the collection $\widehat{\Xi}\left(t\right)$. Then, the conditional expectation in (12.9.1) will be estimated by

$$\widehat{X}\left(t+h;\tau_f,0\right)=\widehat{\Upsilon}\left(t;\tau_f\right)p_f \tag{12.9.12}$$

Similarly, the estimate for the number of males births during the time interval $[t,t+h)$ of type $\mathfrak{t}_m=\left(\tau_m,0\right)$ is

$$\widehat{Y}\left(t+h;\tau_m,0\right)=\widehat{\Upsilon}\left(t;\tau_m\right)p_m. \tag{12.9.13}$$

From now on, let $\widehat{\Xi}\left(t\right)$ be a collection of estimates of the random functions of the process up to time t. In other words, in what follows it will be assumed that these estimates have been made so estimates at time $t+h$ may be calculated recursively. With regard to equation (12.8.2), observe that for $x=1,2,\ldots,r-1$

$$E\left[X_S\left(\left(t;\tau_f,x-1\right)\mid\Xi\left(t\right)\right)\right]=X\left(t;\tau_f,x-1\right)exp\left(-\theta_{tot}\left(t;\tau_f,x-1\right)h\right), \tag{12.9.14}$$

see (12.4.6). Let $\widehat{X}\left(t;\tau_f,x-1\right)$ be an estimate of $X\left(t;\tau_f,x-1\right)$ and let

$$\widehat{\theta}_{tot}\left(t;\tau_f,x-1\right) \tag{12.9.15}$$

be an estimate of $\theta_{tot}\left(t;\tau_f,x-1\right)$, based on estimates of the sample functions of the process at time t. Then, for ages $x=1,2,\ldots,x_m-1$, the estimate of the random function $X\left(t+h;\tau_f,x\right)$ is

$$\begin{aligned}\widehat{X}\left(t+h;\tau_f,x\right)=\widehat{X}_S\left(t;\tau_f,x-1\right)&=\widehat{X}\left(t;\tau_f,x-1\right)\\&\quad\times exp\left(-\widehat{\theta}_{tot}\left(t;\tau_f,x-1\right)h\right),\end{aligned} \tag{12.9.16}$$

see (12.4.6). Similarly, for males at ages $y = 1, 2, \ldots, x_m - 1$, the estimate of the random function $Y(t + h; \tau_m, y)$ is

$$\widehat{Y}(t + h; \tau_m, y) = \widehat{Y}_S(\tau_m, y - 1) = \widehat{Y}(t; \tau_m, y - 1) \times exp\left(-\widehat{\theta}_{tot}(t; \tau_m, y - 1)\, h\right). \tag{12.9.17}$$

With regard to equation (12.8.3), let $\widehat{X}_C(t; \tau_f, x - 1)$ denote an estimate of the random function $X_C(t; \tau_f, x - 1)$. Then from (12.5.22) it can be seen that

$$\widehat{X}_C(t; \tau_f, x) = \sum_{\tau_m \in \mathfrak{T}} \sum_{y = x_m}^{r} \widehat{Z}_C(t; \tau_f, x; \tau_m, y), \tag{12.9.18}$$

where from (12.5.21) it can be seen that

$$\widehat{Z}_C(t; \tau_f, x; \tau_m, y) = \widehat{N}(t; \kappa_c)\, p(\kappa_c) \tag{12.9.19}$$

for $\kappa_c = (\tau_f, x; \tau_m, y)$. The estimate $\widehat{N}(t; \kappa_c)$ would in turn be computed by using the algorithms implicit in equation (12.5.18), and based on estimates of the random functions of the process at time t.

The next step in the derivation of non-linear difference equations embedded in the stochastic population process is that of describing procedures for computing an estimate of the random function $X_{DIS}(t; \tau_f, x - 1)$, see (12.8.3). Since this random function was defined in section 12.6 in connection with the construction of a model governing the evolution of coupe types, the results that follow will also be applicable to estimating the random functions in equation (12.8.7) describing the evolution of couple types. With regard to the vector-valued random function $\boldsymbol{W}(t; \kappa_c)$ in (12.6.8) and vector of conditional probabilities $\boldsymbol{\Pi}(t; \kappa_c)$ in (12.6.9), it can be seen from (12.6.10) that an estimate of the random vector $\boldsymbol{W}(t; \kappa_c)$ is

$$\widehat{\boldsymbol{W}}(t; \kappa_c) = \widehat{Z}(t; \kappa_c) \times \widehat{\boldsymbol{\Pi}}(t; \kappa_c), \tag{12.9.20}$$

where $\widehat{Z}(t; \kappa_c)$ is a scalar and $\widehat{\boldsymbol{\Pi}}(t; \kappa_c)$ is an estimate of the random vectors in (12.6.8) and (12.6.9) based on estimate of the random functions of the process at time t. Therefore, by substituting the components in the vector $\widehat{\boldsymbol{W}}(t; \kappa_c)$ into formulas in section 12.6, estimates of the random functions $X_{DIS}(t; \tau_f, x - 1)$ and $Y_{DIS}(t; \tau_m, x - 1)$ may be computed using the formulas (12.6.10) through (12.6.14). This estimation procedure would lead to a pair of equations of the form

$$\widehat{X}(t + h; \tau_f, x) = \widehat{X}_S(t; \tau_f, x - 1) - \widehat{X}_C(t; \tau_f, x - 1) + \widehat{X}_{DIS}(t; \tau_f, x - 1) \tag{12.9.21}$$

and

$$\widehat{Y}\left(t+h;\tau_m,x\right)=\widehat{Y}_S\left(t;\tau_m,x-1\right)-\widehat{Y}_C\left(t;\tau_m,x-1\right)+\widehat{Y}_{DIS}\left(t;\tau_m,x-1\right), \tag{12.9.22}$$

which are valid for ages $x=x_m,\ldots,r$, single females of type $\mathfrak{t}_f=(\tau_f,x)$ and single males of type $\mathfrak{t}_m=(\tau_m,x)$.

Finally, estimates of the random functions in equation (12.8.7), describing the evolution of couple types, may also be obtained by using the fourth element in the vector $\widehat{\boldsymbol{W}}\left(t;\kappa_c\right)$ and the estimate $\widehat{Z}_C\left(t;\kappa_c\right)$ in (12.9.19). Thus, for $\kappa_c'=(\tau_f,\tau_m;x-1,y-1)$ and $\kappa_c=(\tau_f,\tau_m;x,y)$, it follows from (12.8.7) that

$$\widehat{Z}\left(t+h;\kappa_c\right)=\widehat{Z}_C\left(t;\kappa_c'\right)+\widehat{W}\left(t;\kappa_c',\kappa_c\right), \tag{12.9.23}$$

where $\widehat{W}\left(t;\kappa_c',\kappa_c\right)$ is the fourth element in the vector $\widehat{\boldsymbol{W}}\left(t;\kappa_c\right)$.

12.10 A Two Sex Age Dependent Population Process Without Couple Formation

In the foregoing sections of this chapter, it was observed that when a module for couple formation, depending on the ages of the female and male in a couple, was an integral part of the formulation of an age dependent stochastic population process, then the number of couple types may become very large, which leads to the necessity of processing large arrays in computer implementations of the process. The processing of large arrays in a computer, in turn, becomes problematic, unless some scheme is adopted to reduce the number of couple types. It is of interest, therefore, to formulate a two sex age dependent population process in which sexual contacts may occur between females and males but without partnerships or couples which may last for long time periods. Accordingly, the purpose of this section is to formulate a process that includes sexual contacts between females and males but not partnerships of females and males of long duration. Just as in the foregoing sections of this chapter, only three genotypes with respect to two alleles at some autosomal locus and $r+1$ age classes will be under consideration. Again let $x=x_m,\ldots,x_{\max}$ denote the fertile ages for females and let $y=x_m,\ldots,r$ denote the ages of males such that they may sire offspring. From now on such males will be referred to as fertile.

To reduce the size of the arrays to be processed in a computer, it will be assumed that sexual contacts between females and males do not depend on the age of the male but only on the frequency of the male's genotype. For

$y = x_m, \ldots, r$, let the random function $Y(t; \tau_m, y)$ denote the number of males of type $\mathfrak{t}_m = (\tau_m, y)$ in the population at time t. Then, the random function

$$Y(t; \tau_m) = \sum_{y=x_m}^{r} Y(t; \tau_m, y) \tag{12.10.1}$$

is the total number of fertile males of genotype τ_m in the population at time t. Therefore, the frequency of fertile males of genotype τ_m in the population at time t is given by the random function

$$U_m(t; \tau_m) = \frac{Y(t; \tau_m)}{Y(t; \circ)}, \tag{12.10.2}$$

for $Y(t; \circ) > 0$, where

$$Y(t; \circ) = \sum_{\tau_m} Y(t; \tau_m) \tag{12.10.3}$$

and $U_m(t; \tau_m) = 0$ if $Y(t; \circ) = 0$.

With a view towards minimizing the size of arrays that need to be processed in a computer, it will be assumed that for each fertile female of age x expresses preferences for male sexual partners by genotype. Therefore, just as in chapter 11, denote the three genotypes under consideration by $\tau_1 = AA, \tau_2 = Aa$ and $\tau_3 = aa$ and let

$$\boldsymbol{A}_f = \begin{pmatrix} \alpha_f(1,1) & \alpha_f(1,2) & \alpha_f(1,3) \\ \alpha_f(2,1) & \alpha_f(2,2) & \alpha_f(2,3) \\ \alpha_f(3,1) & \alpha_f(3,2) & \alpha_f(3,3) \end{pmatrix} \tag{12.10.4}$$

denote the 3×3 matrix of acceptance probabilities for females. For example, $\alpha_f(1,1)$ is the conditional probability that a female of genotype τ_1 finds a male of genotype τ_1 acceptable as a sexual partner, and in general, given a female of genotype τ_f, $\alpha_f(\tau_f, \tau_m)$ is the conditional probability that she finds a male of genotype τ_m acceptable as a sexual partner. For the sake of simplicity, it will be assumed that this matrix does not depend on the age x of a fertile female.

Given these definitions, let

$$\gamma_f(t; \tau_f, \tau_m) = \frac{U_m(t; \tau_m)\, \alpha_f(\tau_f, \tau_m)}{\sum_{\tau_m} U_m(t; \tau_m)\, \alpha_f(\tau_f, \tau_m)} \tag{12.10.5}$$

denote the conditional probability that a female of genotype τ_f has a sexual contact with a male of genotype τ_m during the time interval $[t; t+h)$, and let

$$\boldsymbol{\gamma}_f(t; \tau_f) = \left(\gamma_f(t; \tau_f, \tau_\nu) \mid \nu = 1, 2, 3\right) \tag{12.10.6}$$

denote a 1×3 vector of these conditional probabilities. At time t, for $x = x_m, \ldots, x_{\max}$ let the random function $X(t; \tau_f, x)$ denote the number of fertile females of type $\mathfrak{t}_f = (\tau_f, x)$ in the population at time t, and let the random function $Z(t; \mathfrak{t}_f, \tau_m)$ denote the number of females of type $\mathfrak{t}_f = (\tau_f, x)$ who have a sexual contacts with a males of genotype τ_m during the time interval $[t, t+h)$. Then, let

$$\boldsymbol{Z}(t; \mathfrak{t}_f) = (Z(t; \mathfrak{t}_f, \tau_{m_\nu}) \mid \nu = 1, 2, 3) \tag{12.10.7}$$

denote a 1×3 random vector whose elements are the indicated random functions. It will be assumed that this random vector has a conditional multinomial distribution with index $X(t; \tau_f, x)$ and probability vector $\boldsymbol{\gamma}_f(t; \tau_f)$. In symbols,

$$\boldsymbol{Z}(t; \mathfrak{t}_f) \sim CMultinom\left(X(t; \tau_f, x), \boldsymbol{\gamma}_f(t; \tau_f)\right). \tag{12.10.8}$$

Reproductive success for each fertile female who has sexual contacts with a fertile male, will be characterized by the parameters $\lambda(\tau_f, x)$ denoting the expected number of offspring each fertile female type $\mathfrak{t}_f = (\tau_f, x)$ contributes to the population during a time interval $[t, t+h)$. More precisely, if $N(\tau_f, x)$ a random variable taking values in the set of nonnegative integers, then it will be assumed that $N(\tau_f, x)$ has a Poisson distribution with parameter $\lambda(\tau_f, x)$. Given $Z(t; \mathfrak{t}_f, \tau_{m_\nu})$, let $N_{v'}(\tau_f, x)$ for $\nu' = 1, 2, \ldots, Z(t; \mathfrak{t}_f, \tau_{m_\nu})$ denote a collection of conditionally independent random variables such that the distribution of each of these random variables is that of $N(\tau_f, x)$. Then the total number of offspring arising from sexual contact of type $\kappa_s = (\mathfrak{t}_f, \tau_{m_\nu})$ between females of type $\mathfrak{t}_f = (\tau_f, x)$ and males of genotype τ_{m_ν} during the time interval $[t, t+h)$ is given by the random variable

$$H(t; \kappa_s) = \sum_{\nu'=1}^{Z(t; \mathfrak{t}_f, \tau_{m_\nu})} N_{v'}(\tau_f, x) \tag{12.10.9}$$

for $Z(t; \mathfrak{t}_f, \tau_{m_\nu}) > 0$ and $H(t; \kappa_s) = 0$ if $Z(t; \mathfrak{t}_f, \tau_{m_\nu}) = 0$. Observe that, given $Z(t; \mathfrak{t}_f, \tau_{m_\nu})$, the random function $H(t; \kappa_s)$ has a Poisson distribution with parameter $Z(t; \mathfrak{t}_f, \tau_{m_\nu}) \lambda(\tau_f, x)$.

Let $\boldsymbol{p}(\tau_f, \tau_m; \boldsymbol{\tau})$ denote the 1×3 vector in (12.7.6) and let the 1×3 random vector

$$\boldsymbol{V}(t; \tau_f, \tau_m; x, \boldsymbol{\tau}) = (V(t; \tau_f, \tau_m; x, \circ; \tau_\nu) \mid \nu = 1, 2, 3) \tag{12.10.10}$$

denote the number of offspring of each of the three genotypes produced by females with τ_f of age x at time t from sexual contacts with males of genotype τ_m during the time interval $[t, t+h)$. Then, by assumption,

$$\boldsymbol{V}(t; \tau_f, \tau_m; x, \boldsymbol{\tau}) \sim CMultinom\left(H(t; \kappa_s), \boldsymbol{p}(\tau_f, \tau_m; \boldsymbol{\tau})\right). \tag{12.10.11}$$

For a given type of sexual contact $\kappa_s = (\tau_f, \tau_m)$, the 1×3 random vector

$$\boldsymbol{V}\left(t; \tau_f, \tau_m; \circ, \boldsymbol{\tau}\right) = \sum_{x=x_m}^{x_{\max}} \boldsymbol{V}\left(t; \tau_f, \tau_m; x, \boldsymbol{\tau}\right) \tag{12.10.12}$$

contains the number of offspring of each of the three genotypes produced from sexual contacts of type $\kappa_s = (\tau_f, \tau_m)$ during the time interval $[t, t+h)$. Therefore the random vector

$$\boldsymbol{V}\left(t; \boldsymbol{\tau}\right) = \sum_{\nu=1}^{3} \sum_{\nu'=1}^{3} \boldsymbol{V}\left(t; \tau_{f_v}, \tau_{m_{\nu'}}; \circ, \boldsymbol{\tau}\right) \tag{12.10.13}$$

contains the number of each of the three genotypes produced from all sexual contacts during the time interval $[t, t+h)$.

Let $V\left(t; \boldsymbol{\tau}_v\right)$ for $\nu = 1, 2, 3$ denote the three random components of the vector $\boldsymbol{V}\left(t; \boldsymbol{\tau}\right)$. Then, the number of females of genotype τ_ν born during the time interval $[t, t+h)$ is given by the random function

$$B_f\left(t; \tau_\nu, 0\right) \sim CBinom\left(V\left(t; \boldsymbol{\tau}_v\right), p_f\right), \tag{12.10.14}$$

and the number of males with this genotype born during this time interval is

$$B_m\left(t; \tau_\nu, 0\right) = V\left(t; \boldsymbol{\tau}_v\right) - B_f\left(t; \tau_\nu, 0\right) \tag{12.10.15}$$

for $\nu = 1, 2, 3$. Let the random function $X\left(t+h; \tau_f, 0\right)$ denote the number of females of type $\mathfrak{t}_f = (\tau_f, 0)$ in the population at time $t+h$, and define the random function $Y\left(t+h; \tau_m, 0\right)$ similarly for males. Then,

$$X\left(t+h; \tau_f, 0\right) = B_f\left(t; \tau_f, 0\right) \tag{12.10.16}$$

and

$$Y\left(t+h; \tau_m, 0\right) = B_m\left(t; \tau_m, 0\right) \tag{12.10.17}$$

for all genotype τ_f and τ_m.

In general, let the random functions $X\left(t+h; \tau_f, x\right)$ and $Y\left(t+h; \tau_m, y\right)$ denote, respectively, the number of females of type $\mathfrak{t}_f = (\tau_f, x)$ and the number of males of type $\mathfrak{t}_m = (\tau_m, y)$ in the population at time $t+h$. Then, the algorithms for computing realizations of these random functions from their values $X\left(t; \tau_f, x-1\right)$ and $Y\left(t; \tau_m, y-1\right)$ at time t for ages $x, y \geq 1$ are the same as those outlined in section 12.8 for single females and males, see, for example, (12.4.6) for more details. Similarly, the procedures for computing the deterministic estimates $\widehat{X}\left(t+h; \tau_f, x\right)$ and $\widehat{Y}\left(t+h; \tau_m, y\right)$ of these random functions follows the algorithms set forth in section 12.9. Furthermore, by using conditional expectations as outlined in section 12.9

and the Monte Carlo simulation algorithms described in this section, deterministic estimates $\widehat{B}_f\left(t;\tau_f,0\right)$ and $\widehat{B}_m\left(t;\tau_m,0\right)$ of the random functions in (12.10.16) and (112.10.17) could be computed. An advantage of the formulation outlined in this section is that it requires the processing of relatively small arrays in a computer when compared with the formulation which involved the formation and evolution of couples described in the foregoing sections. The processing of smaller arrays, in turn, will result in shorter times for computers to complete Monte Carlo simulation experiments, which will increase the feasibility of doing exploratory simulation experiments with age dependent population processes.

12.11 Parametric Latent Risk Functions for Death by Age

As was shown in the foregoing sections of this chapter, latent risk functions for death by age play an essential role in formulating stochastic versions of age dependent models for the evolution of populations with two sexes. Parametric forms of latent risk have many advantages when designing software to implement age dependent models. For if estimates are the parameters are available or values of these parameters may be assinged through a process of plausible reasoning, then numerical arrays computed form the latent risk functions may be generated in a computer with relative ease. Accordingly, the purpose of this section is to provide examples of parametric risk functions that have been useful in the study of human and animal mortality. The basic strategy to be followed in this section is that of partitioning the life span of an individual into two stages consisting of sexually immature individuals and sexually mature adults.

For many species of animals, offspring are at high risk of death following hatching or birth but as time passes the risk of death decreases. Therefore, the latent risk function for infant deaths and sexually immature individuals will be assumed to have the exponential form

$$\theta_0(x)=\alpha_0\beta_0\exp\left[-\beta_0 x\right], \tag{12.11.1}$$

where $x=0,1,\ldots,x_m$ and α_0 and β_0 are positive parameters. As can be seen form this formula, the latent risk of death for infants born at time t during the time interval $[t,t+h)$ is the constant $\theta_0(0)=\alpha_0\beta_0$. The integral of this latent risk function is

$$H_0(x)=\int_0^x\theta_0(s)ds=\alpha_0\left(1-\exp\left[-\beta_0 x\right]\right) \tag{12.11.2}$$

for $x \geq 0$, and, by definition, the latent survival function corresponding to this risk function is

$$S_0(x) = \exp\left[-H_0(x)\right]. \tag{12.11.3}$$

In terms of this component of the model, the probability an individual survives infancy is the limit

$$\lim_{x \uparrow \infty} S_0(x) = S_0 = e^{-\alpha_0}. \tag{12.11.4}$$

Quite often an investigator will have some knowledge of the fraction S_0, which may be used to obtain an initial or trial estimate of α_0; namely $\alpha_0 = -\ln S_0$. In humans, infant boys have a greater risk of death than girls, and moreover, these risks may depend on the genotype of the infant male or female. To accommodate such dependence the parameter α_0 will be denoted as a function of sex and genotype by the symbols $\alpha_0\left(\tau_f\right)$ and $\alpha_0\left(\tau_m\right)$.

Next observe that the term in parenthesis

$$F_0(x) = 1 - \exp\left[-\beta_0 x\right] \tag{12.11.5}$$

in (12.11.2) is the distribution function of a random variable X_0 with expectation

$$E\left[X_0\right] = \frac{1}{\beta_0}. \tag{12.11.6}$$

Thus, the parameter β_0 will determine the speed at which deaths occur. Small values of β_0 will correspond to longer infant survival times; while large values of β_0 will correspond to shorter survival times. Such ideas may be quantified by assigning trial values to $E\left[X_0\right]$ to find an estimate of the form $\beta_0 = 1/E\left[X_0\right]$. For example, a plausible value of this expectation may be chosen as $E\left[X_0\right] = x_m/2$, which yields $\beta_0 = 2/x_m$ as a preliminary estimate of β_0. The parameter β_0 may also depend on the sex and genotype of an individual and will be denoted by the symbols $\beta_0\left(\tau_f\right)$ and $\beta_0\left(\tau_m\right)$.

To accommodate accidents that may occur throughout the life span of an individual, the latent risk function for this component will have the simple form $\theta_1(x) = \alpha_1$ for all $x \geq 0$, where α_1 is a positive constant. In this case, the integral of the risk function is

$$H_1(x) = \int_0^x \theta_1(s)ds = \alpha_1 x \tag{12.11.7}$$

for $x \geq 0$. Therefore, for $x \geq 0$, the latent survival function has the simple form

$$S_1(x) = \exp\left[-\alpha_1 x\right]. \tag{12.11.8}$$

A useful trial value of α_2, based on period studies of human mortality, is about 0.001. This component of the model is often referred to as the Makeham component.

A third two-parameter latent risk function, due to Gompertz (19-th Century), deals with risks of deaths at the older ages. Let α_2 and β_2 be positive parameters. Then, it will be assumed that for $x \geq 0$ the latent risk function $\theta_2(x)$ has the form

$$\theta_2(x) = \alpha_2 \beta_2 \exp\left[\beta_2 x\right]. \tag{12.11.9}$$

Observe that, as it should, this risk function increases as age x of an individual increases, and, by assumption, the risk of death increases exponentially with increasing age. The integral of this risk function has the form

$$H_2(x) = \int_0^x \theta_2(s)ds = \alpha_2 \left(\exp\left[\beta_2 x\right] - 1\right) \tag{12.11.10}$$

for $x \geq 0$. Therefore, the latent survival function for this component is

$$S_2(x) = \exp\left[-\alpha_2 \left(\exp\left[\beta_2 x\right] - 1\right)\right] \tag{12.11.11}$$

for $x \geq 0$. By applying a general but standard formula that a density is the risk function times the survival function, it can be seen that the probability density function of the Gompertz distribution has the form

$$f_2(x) = \theta_2(x)S_2(x) = \alpha_2\beta_2 \exp\left[\beta_2 x\right] \exp\left[-\alpha_2 \left(\exp\left[\beta_2 x\right] - 1\right)\right] \tag{12.11.12}$$

for $x \geq 0$. Although this distribution may be derived from intuitively appealing assumptions, it is more difficult to handle from a mathematical point of view than some other distributions that arise in probability and statistics. Nevertheless, because many advanced mathematical functions are now available to research workers in such software packages as MAPLE, MATHEMATICA., and MATLAB, an outline of the mathematics used in analyzing the Gompertz distribution seems appropriate.

As the parameters α_2 and β_2 do not have obvious statistical interpretations, such as an expectation or variance, it is difficult to assign tentative values to them. Quite often, however, there is some feeling about the modal age of death for those who survive to old age. Let m_2 denote the mode of the Gompertz distribution. Then by using elementary calculus to find the maximum of the density $f_2(x)$ in (12.11.12), it can be shown that the equation

$$\alpha_2 = \exp\left[-\beta_2 m_2\right] \tag{12.11.13}$$

formalizes a connection among the parameters α_2, β_2 and m_2. In particular, if m_2 is assigned a value and β_2 is known, then α_2 is determined. But, to find a plausible value of β_2, more input is needed.

Let X_2 denote a random variable with a Gompertz distribution. Then, after considerable analysis, it can be shown that the exact formula for the expectation is

$$E\left[X_2\right] = e^{\alpha_2}\left[m_2 - \frac{C}{\beta_2} + \frac{1}{\beta_2}\sum_{\upsilon=0}^{\infty}\frac{(-1)^{\nu}\alpha_2^{\nu+1}}{\nu!(\nu+1)^2}\right], \qquad (12.11.14)$$

where $C \simeq 0.57721\cdots$ is Euler's constant. Furthermore, the exact formula for the second moment is

$$E\left[X_2^2\right] = e^{\alpha_2}\left[\left(m_2 - \frac{C}{\beta_2}\right)^2 + \frac{\pi^2}{\beta_2^2 6} - \frac{2}{\beta_2^2}\sum_{\upsilon=0}^{\infty}\frac{(-1)^{\nu}\alpha_2^{\nu+1}}{\nu!(\nu+1)^3}\right]. \qquad (12.11.15)$$

.

When $\alpha_2 > 0$ is small, then $\exp[\alpha_2] \simeq 1$ and the above infinite series may be neglected. Thus, the approximations

$$E\left[X_2\right] \simeq m_2 - \frac{C}{\beta_2} \qquad (12.11.16)$$

and

$$E\left[X_2^2\right] \simeq \left(m_2 - \frac{C}{\beta_2}\right)^2 + \frac{\pi^2}{\beta_2^2 6} \qquad (11.11.17)$$

hold for small α_2. Therefore, a formula for the approximate variance of the Gompertz distribution is

$$\sigma_2^2 \simeq \frac{\pi^2}{\beta_2^2 6}. \qquad (12.11.18)$$

Equivalently,

$$\beta_2 \simeq \frac{\pi}{\sigma_2\sqrt{6}}. \qquad (12.11.19)$$

It will be instructive to present a simple numerical example, illustrating the use of these approximations. Suppose, for example, one wished to construct a Gompertz distribution with a mode $m_2 = 60$ years and suppose the standard deviation of this distribution is $\sigma_2 = 10$. Then, formula (12.11.19) yields the approximation $\beta_2 \simeq 0.128254983\,016\,86$. Furthermore, by using (12.11.13), it can be seen that this formula yields the approximation $\alpha_2 = 4.549609\,43585548 \times 10^{-4}$. Moreover,

$$\exp\left(4.\,549609435\,855\,48 \times 10^{-4}\right) = 1.000455064454\,01 \simeq 1. \qquad (12.11.20)$$

Due to the approximation in (12.11.19), it appears that the parameter β_2 will belong to the interval $(0, 1)$ for plausible values of σ_2 so that from (12.11.13) it can be seen that when the mode m_2 of the distribution is sufficiently large, the parameter $\alpha_2 > 0$ will be small. As can be seen form the above discussion, when α_2 is small, the approximation in (12.11.19) will yield good results. A more extensive account of parametric forms of risk functions for death has be given in Mode and Sleeman (2000) section 13.2.

To expedite the computer implementation of the results presented in this section, the intrinsic risk function $\theta_{intrin}\left(\tau_f, x\right)$ for females of genotype τ_f and age x defined section 12.4 will be expressed in terms of the parametric risk functions defined in this section. For example, if it is assumed that the risk function in (12.1.1) and the Makeham constant $\alpha_1\left(\tau_f\right)$ applies for ages $x = 0, 1, 2, \ldots, x_m$, then

$$\theta_{intrin}\left(\tau_f, x\right) = \alpha_0\left(\tau_f\right)\beta_0\left(\tau_f\right)\exp\left[-\beta_0\left(\tau_f\right)x\right] + \alpha_1\left(\tau_f\right). \tag{12.11.21}$$

Therefore, the total risk function defined in (12.4.3) becomes

$$\begin{aligned}\theta_{tot}\left(t; \tau_f, x\right) = {} & \alpha_0\left(\tau_f\right)\beta_0\left(\tau_f\right)\exp\left[-\beta_0\left(\tau_f\right)x\right] + \alpha_1\left(\tau_f\right) \\ & + \alpha\left(\tau_f\right)\beta^{\alpha\left(\tau_f\right)}\left(\tau_f\right)\left(Z_{tot}\left(t\right)\right)^{\alpha\left(\tau_f\right)-1},\end{aligned} \tag{12.11.22}$$

where $Z_{tot}\left(t\right)$ is total population size at time t.

The formula for computing a realization of the random function $Z_{tot}\left(t\right)$ will depend on the class of stochastic process under consideration. For example, when the formulation with no couple formation described in section 12.10 is under consideration, then this random would be computed using the following expressions. As in that section, let $X\left(t; \tau_f, x\right)$ denote the number of females of type $\mathfrak{t}_f = \left(\tau_f, x\right)$ in the population at time t and let the random function $Y\left(t; \tau_m, y\right)$ be defined similarly for males. Then, let the random function

$$X\left(t; \circ, \circ\right) = \sum_{\tau_f \in \mathfrak{T}} \sum_{x=0}^{r} X\left(t; \tau_f, x\right) \tag{12.11.23}$$

denote the total number of females in the population at time t, and let $Y\left(t; \circ, \circ\right)$ denote the corresponding random function for males. Then, $Z_{tot}\left(t\right) = X\left(t; \circ, \circ\right) + Y\left(t; \circ, \circ\right)$. A derivation of the formula for computing $Z_{tot}\left(t\right)$ for processes in which couple formation occurs will be left as an exercise for the reader.

For ages $x = x_m + 1, \ldots, r$ it will be assumed that the Gompertz risk function in (12.12.9) is in force so that intrinsic risk function $\theta_{intrin}\left(\tau_f, x\right)$ takes the form

$$\theta_{intrin}\left(\tau_f, x\right) = \alpha_2\left(\tau_f\right)\beta_2\left(\tau_f\right)\exp\left[\beta_2\left(\tau_f\right)x\right] + \alpha_1\left(\tau_f\right). \tag{12.11.24}$$

Thus, for these ages the total risk function has the form

$$\theta_{tot}\left(t;\tau_f,x\right)=\alpha_2\left(\tau_f\right)\beta_2\left(\tau_f\right)\exp\left[\beta_2\left(\tau_f\right)x\right]\ +\alpha_1\left(\tau_f\right)$$
$$+\alpha\left(\tau_f\right)\beta^{\alpha(\tau_f)}\left(\tau_f\right)\left(Z_{tot}\left(t\right)\right)^{\alpha(\tau_f)-1}. \quad (12.11.25)$$

Similar formulas for the total risk functions for males of type $\mathfrak{t}_m=(\tau_m,y)$ may also be set down, but these formulas would be identical in form to those in this section except the parameters for males may differ from those for females.

12.12 Sexual Selection in an Age Dependent Process Without Couple Formation

In this section, sexual selection will be studied in a computer experiment based on the two sex age dependent process without couple formation described in section 12.10. As a Monte Carlo simulation experiment based on a stochastic age dependent process with 50 to 100 age classes would require a long period of computer time to complete an experiment, attention was will be confined to computer experiments based on the deterministic model embedded in the stochastic process, which were completed in reasonably short periods of time. At the outset, it should also be stated that only one autosomal locus is under consideration with two alleles denoted by A and a. Furthermore, it will be assumed that all individuals in the initial population are of genotype AA and the allele a arises in the population by A mutating to a. In what follows, the assignments of numerical values refer to those parameters defined in sections 12.10 and 12.11. Unlike the two sex process studied in chapter 11, where evolutionary time was expressed in terms of generation, in this chapter time is expressed in terms of years, which would be useful, for example, when one is thinking about evolution that has occurred since the end of the ice age about 10,000 years ago.

The maximum age of any individual in the population in experiment 12.12.1 was assigned the value $r=100$ years for both females and males and it was assumed that the age of sexual maturity for both sexes was $x_m=15$ years. The greatest age $x_{\max}$ at which females were capable of bearing an offspring was assigned the value $x_{\max}=50$. A 3×3 matrix of acceptance probabilities for females was chosen as

$$\boldsymbol{A}_f=\begin{pmatrix}0.1 & 0.1 & 0.9\\ 0.1 & 0.1 & 0.9\\ 0.1 & 0.1 & 0.9\end{pmatrix}, \quad (12.12.1)$$

which is the same set of values used in the experiment on sexual selection reported in section 11.11. As a reminder of the interpretation of this matrix, the three genotypes under consideration, AA, Aa, aa, are numbered as 1,2, and 3. Thus, according to the matrix in (12.12.1), females of all genotypes prefer males of genotype aa as sexual partners. Implicit in the assignment of the parameters in the matrix is the assumption that allele A is dominant to allele a. By assumption, according to the model presented in section 12.10, males have sexual contacts with females at random in the sense that the probabilities of these contacts depend only on the frequency of each of the three male genotypes in the male population and not on age as described in section 12.10.

The expected number of potential offspring produced by females of all genotypes during their life span was assinged the value 12, indicating that, by assumption, selection was neutral with respect reproductive success of the females of each of the three genotypes. To accommodate the effect of age on reproduction, the value 12 was multiplied by an array of Poisson probabilities that had the property that the modal age class for births was age class 25 for females, which reflects the idea that the mid-twenties are the optimal ages for human females to bear children. The reason for assigning the values 12 was the desire to simulate the rapid evolution of a population founded by a relatively small number of individuals. The matrix of mutation probabilities chosen for this experiment was

$$\mathfrak{M} = \begin{pmatrix} \mu_{11} & 10^{-5} \\ 10^{-6} & \mu_{22} \end{pmatrix}, \tag{12.12.1}$$

where $\mu_{12} = 10^{-5}$ is the probability of mutation $A \rightarrow a$ and $\mu_{21} = 10^{-6}$ is the probability of the mutation $a \rightarrow A$. These numbers will be interpreted as the probabilities of observing these mutations during the process of gametogenesis for both females and males in the reproductive ages. With reference to the assignment of values for parameters for the Weibull type risk functions that take into account density dependence, the α-parameters for the risk functions for each of the three genotypes were assigned the constant value $\alpha = 2$. Similarly, the β-parameter for each genotype, which determines the ultimate size of the population, were assigned the constant value $\beta = 10^{-7}$. Therefore, given these assignments of parameter values, there was no selection among the three genotypes in their abilities to compete for resources. Observe that, by assumption, these assignments of parameter values apply to both females and males in the population.

With regard to survival following birth, the α_0-parameters for the latent risk function in (12.11.2) were computed by assuming that the probability each infant survives to sexual maturity was 0.9 for all genotypes and both sexes. Given this assumption, the formula that was used to calculate the α_0 parameter for all genotypes and both sexes was that determined by equation (12.11.4).

As was suggested in section 12.11, the β_0 parameter for all genotypes and sexes was chosen as $\beta_0 = 2/x_m$. Furthermore, the Makeham parameter α_1 for all genotypes and both sexes was assigned the value $\alpha_1 = 0.001$. From these parameter assignments, it can be seen that, by assumption, there was no selection with respect to survival in early life or to death from accidents throughout life among the three genotypes and two sexes.

The last two parameters regarding the survival of individuals after they reach maturity were those for the Gompertz latent risk function discussed in section 12.11. As was shown in this section, the parameters α_2 and β_2 in this risk function may be determined by specifying the mode and standard deviation of the Gompertz distribution and using formulas (12.11.13) and (12.11.19). Thus, for all genotypes of both sexes, if the standard deviation were assinged the value $\sigma = 10$, then a value of the β_2 parameter was determined by using the approximation in (12.11.19). Then, by assuming that the mode of the Gompertz distribution was $m_2 = 60$, the α_2 parameter for all genotypes and both sexes was determined by using equation (12.11.13). The last parameter to be assinged a value was that for the probability that an infant is female at birth. As was the case for all computer experiments with two sex models reported in this book, this probability was assinged the value $p_f = 100/205$.

The last step in the development of computer input for the age dependent two sex formulation under consideration is that assigning numbers to each of the $r + 1$ age classes for all genotypes and both sexes. Under the assumption that the only genotype present in the initial population is AA, the problem of assigning these numbers reduces to assigning number of individuals for each of the $r + 1$ age classes for this genotype and for both sexes. In the computer experiment under consideration, it was assumed that the number of individuals of genotype AA in each of the $r + 1 = 101$ age classes was 100 for both sexes. Thus, the total number of individuals in the initial population for each sex in experiment 12.12.1 was $101 \times 100 = 10,100$. The assumption that the age distribution for each sex was uniform in the initial population was, of course, unrealistic, but, as is known form previous experience in projecting age dependent population on a time scale of one year, the effects of the initial age distributions disappear within a few hundred years. Thus, in evolutionary studies where projections cover time periods of at least 6,000 years, the assumed initial age distributions are of little consequence as will be shown subsequently in graphs of the age distribution at chosen years in a projection.

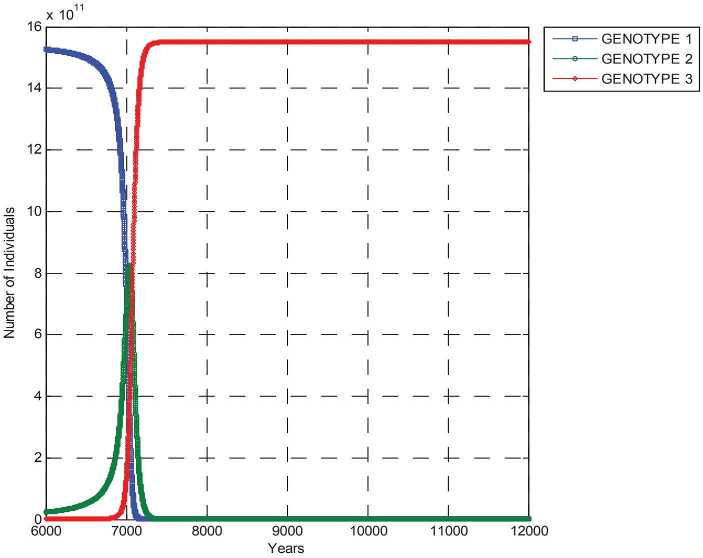

Figure 12.12.1 Trajectories of the Number of Individuals of Each Genotype in the Female Population for Experiment 12.12.1 for Years 6,001 to 12,000.

In experiment 12.12.1 and other experiments reported in subsequent sections, one or more projections were made in blocks of 6,000 years, and, if, for example, the results of a projection were not informative after 6,000 years of evolution, the age distributions for each of the three genotypes for both sexes at 6,000 years were used as the initial age distributions for the next 6,000 years of evolution. Furthermore, this process of computing 6,000 years of evolution was continued until informative results were obtained. On the desk top computer used in these experiments, each 6,000 year segment of an experiment required about 4 hours of computer time, which was an acceptable execution time for experiments that may be continued for several blocks of 6,000 years. In summary, among the four components of natural selection under consideration, preferences of females for males of a given genotype as sexual partners over other males, reproductive success, competitive ability and survival in early and later stages of life, all were neutral except for female preferences for sexual partners.

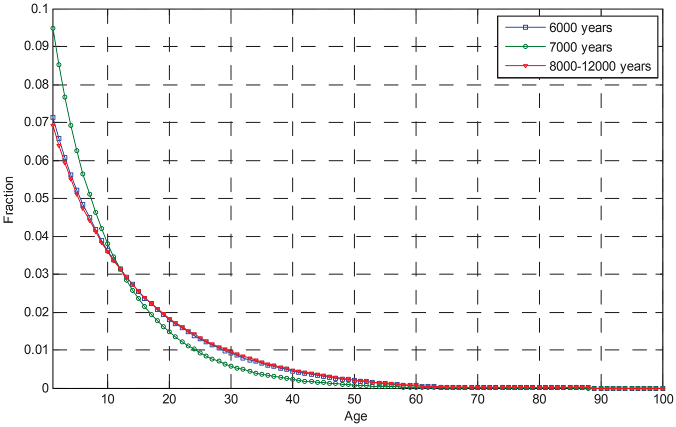

Figure 12.12.2 Graphs of the Age Distribution for Genotype 3, aa, at Selected Years of the Projection in Experiment 12.12.1.

Among the many numerical outputs of the embedded deterministic model was an estimate of the total number of individuals of each genotype in the population for both sexes and at every year in a projection. In experiment 12.12.1, this computer output was examined for the first 6,000 years of the projection but the simulated data were not informative regarding the effects of sexual selection. A decision was made, therefore, to continue the simulation for another 6,000 years. Presented in Figure 12.12.1 are the graphs of the trajectories of the three genotypes for the female population. Although it is not shown, the graphs for the trajectories of the three genotypes for males were very similar. As can be seen from the figure, at about 7,000 years into the projection there is a drastic change in the trajectories of the three genotypes. For example, the trajectory of genotype 1, AA, begins to decline rapidly in terms of evolutionary time along with a rise in the trajectory for genotype 2, Aa, before it undergoes a steep decline to very small numbers a little after 7,000 years into the projection. Concomitantly, there is a steep rise in the trajectory for genotype 3, aa, due to sexual preference of females which culminates with a population size for this genotype at somewhat less than 16×10^{11} individuals. Observe that this graph is similar to those presented in section 11.11, where evolutionary time was expressed in terms for generations.

When dealing an age dependent process, a computer output of interest is the age distribution for each of the three genotypes under consideration at chosen year for both sexes. Shown in Figure 12.12.2 are the graphs of the age distribution for genotype 3, aa, for the female population at selected years. From this figure it can be seen that the age distribution for genotype 3 at 7,000 years into the projection differs from the others shown in the figure. This perturbation of the age distribution is due to the insertion of a large number of individuals of genotype 3 into the population at about 7,000 years into the projection as can be seen in Figure 12.12.1. After this perturbation, however, the age distribution converges to a distribution similar to that of year 6,001 into the projection. Even though the formalities of the stable age distribution have not been treated in this chapter, the shape of the age distribution for the years 8,000 to 12,000 years is consistent with what one would expect from age dependent stable population theory.

12.13 Population Momentum and Emergence of a Beneficial Mutation

In this section some additional results from experiments with the two sex age dependent model without couple formation introduced in section 12.10 will be reported. In experiment 12.12.2, all the parameters assignments were the same as those in experiment 12.12.1 except that in the matrix $\boldsymbol{A}_f$ of acceptance probabilities for females all elements were assigned the number 1, indicating that females did not show preferences for sexual partners among the three genotypes in the male population. In other words, in this experiment matings among females and males were random. Just as in experiment 12.12.1, the initial population was composed of individuals of genotype 1, AA, so that individuals with genotypes Aa and aa would arise in the population only from the process of mutation. The initial age distribution for this experiment was also that used in experiment 12.12.1. To test whether differences among the three females genotypes with respect to reproductive success were sufficient for the emergence and eventual predominance of the genotype aa, the expected number of offspring contributed of the population during their fertile period by each of the three genotypes in the female population were assigned the values $\boldsymbol{\lambda} = (12, 12, 12.5)$. Thus, according this assignment, females of genotype 3, aa, would on average contribute 12.5 offspring to the population during their fertile years and genotypes 1 and 2, AA and Aa, would contribute about 12 offspring to the population during these years. As by assumption, genotype 3 has a selective advantage over others, one would expect in the long run that individuals of genotype 3 would eventually become predominant in the population. However, even after 24,000 years of simulated evolution, individuals of genotype 3 both sexes were present in the population only in small numbers when compared with individuals of the initial genotype 1.

In experiment 12.12.3, all the parameter assignments as well as the initial age distributions were the same as in experiment 12.12.2 except that it was assumed that the expected numbers of offspring produced by females during their fertile periods by each the three genotype were as indicated in the vector $\boldsymbol{\lambda} = (6, 6, 12)$. Thus, in this experiment, it was assumed that individuals of genotype aa had a significant selective advantage over the other two genotypes due to greater reproductive success.

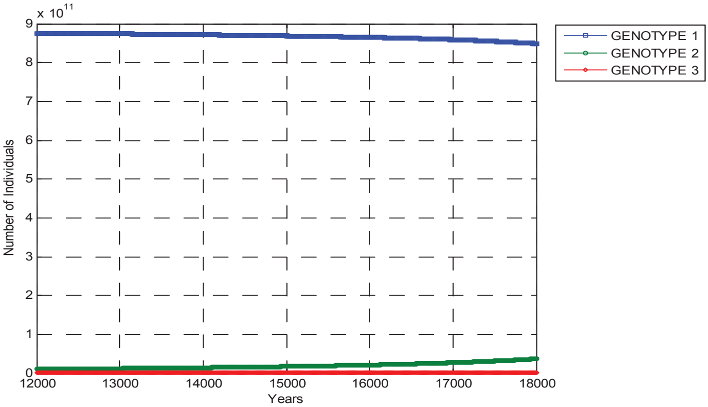

Figure 12.13.1 Graphs of the Trajectories for the Total Numbers of the Three Genotypes in the Female Population for the Years 12,001 to 18,000 for Experiment 12.12.3.

Presented in Figure 12.13.1 are the graphs of the trajectories for the total numbers of the three genotypes in the female population during years 12,001 to 18,000 of the projection as determined by the embedded deterministic model. From this figure, it can be seen that even by 18,000 years into the projection, the numbers of individual carrying the mutation allele a are rather small when compared with the initial genotype AA that made up the initial population.

Even though there is some suggestion that the number of individuals in the population of genotype 1 are declining, the rise to predominance of the genotype aa is very slow so that it would require many more years of evolution than 18,000 before this genotype became predominant in the population. In this experiment population momentum, characterized by the predominance of genotype AA in a population of slowly changing total population size, significantly delayed the rise in predominance of genotype aa. The graph trajectories of the three genotypes in for the male population were similar, and were, therefore, omitted.

In experiment 12.12.4 all the parameter values were the same those for experiment 12.12.3 except that the β-parameters in the Weibull survival function were changed so that genotype aa would have a competitive advantage over genotypes AA and Aa. In this experiment, the vector of $\boldsymbol{\beta}$-parameters was assigned the value $\boldsymbol{\beta} = \left(10^{-6}, 10^{-6}, 10^{-8}\right)$ for both sexes. Thus, by assumption, genotype 3, aa, had a two orders of magnitude competitive advantage over genotypes 1 and 2 in experiment 12.12.4. Presented in Figure 12.13.2 are graphs of the trajectories for the total numbers of the three genotypes as computed from the embedded deterministic model for the female population during the years 6,001 to 12,000 of the projection.

As can be seen from this graph, during the years of the projection from 6,000 to 8,000, the three genotypes were present in the population in approximately equal numbers. On the graph, these number appear small, but, since the scale of the vertical axis is in units of 10^{13}, even a number of the fractional form 0.5×10^{13} would be a large number. This apparent demographic equilibrium continued up to nearly 8,500 years when it was punctuated by a rapid increase in the number of individuals of genotype 3 which reached a value somewhat less than 16×10^{13}.

This rapid rise in the number of individuals of genotype 3 lead to another demographic equilibrium at about 9,000 years in which genotype 3 was predominant in the population and genotypes 1 and 2 were present at much smaller numbers. Some of the properties of the graphs in Figure 12.13.1 are similar to those observed in experiment 12.12.1 in that it appears that the number of individuals of genotype 3 must reach some threshold before there is a rapid rise in the number of individuals of this genotype.

From Figure 12.13.2, it can also be seen that by assigning the values $\boldsymbol{\beta} = \left(10^{-6}, 10^{-6}, 10^{-8}\right)$ for the β-parameters, which resulted in individuals of genotype aa having a greater competitive ability over genotypes 1 and 2, was a decisive factor in the rise of genotype 3 to predominance in the population.

When working with an age dependent models of evolution, displaying changes in the age distribution for a given sex and genotype at selected times in a simulation experiment are always of interest. Contained in Figure

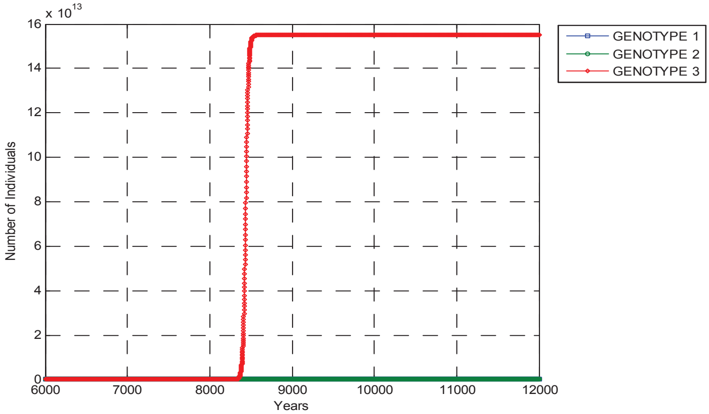

Figure 12.13.2 Graphs of the Trajectories for the Total Numbers of the Three Genotypes in the Female Population for the Years 6,001 to 12,000 for Experiment 12.12.4.

12.13.3 are graphs of the age distribution for the females of genotypes 3 for the years 8,000 and 10,000 for the projection in experiment 12.12.4. As can be seen from the graph for the year 10,000, the age distribution for females of genotype 3 in a monotone decreasing function of age. This shape of an age distribution is typical for an evolving age dependent population in demographic equilibrium. On the other hand, it can also be seen from the Figure that the age distribution for the year 8,000 is not a monotone decreasing function of age, but has a bulge starting at about age 20.

Such bulges in the age distribution are typical for an evolving age dependent population that has under gone a surge in births in the recent past. In experiment 12.12.4, this surge of births in the number of females of genotype 3 near 8,000 years into projection, but these number did no begin their steep assent until about 8,500 years into the projection. In this connection, it is also of interest to remark that, although it is not shown in the Figure 12.13.3, the age distribution for the year 8,700 was identical with that for the years 10,000 as shown in the Figure, indicating that the sub-population of females of genotype 3 reached a demographic equilibrium shortly after population size attained a value somewhat less than 16×10^{13}, see Figure 12.13.2.

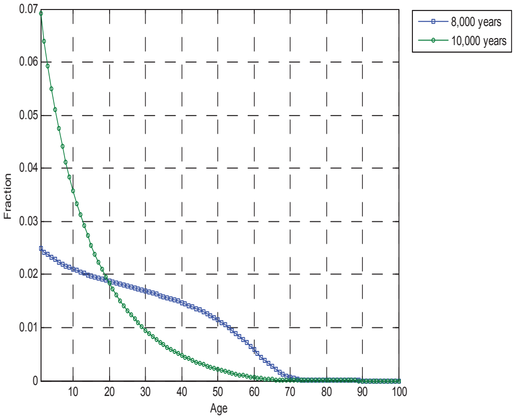

Figure 12.13.3 Graphs of the Age Distributions for Females of Genotype 3 at Years 8,000 and 10,000 for Experiment 12.12.4.

12.14 Experiments with a Version of the Age Dependent Model with Couple Formation

As mentioned in the foregoing sections of this chapter, when a two sex age dependent model with couple formation is considered with 50 or more age classes for both females and males, the number of couple types that arise in a formulation may be very large so that writing software for desk top computers becomes problematic, due to problems of repeatedly processing large arrays in computer experiments involving many years or other time units of evolution. To avoid such problems in a preliminary computer implementation of the class of stochastic processes under consideration, the highest age considered was $r = 12$ with $x_m = 2$ as the age of sexual maturity and $x_{\max} = 6$ as the greatest age of reproduction for both sexes. Under the assumption that only three genotypes with respect to some autosomal locus were under consideration, the number of types of single females and males there were eligible to form couples and reproduce were $3 \times 5 = 15$ for both sexes. Therefore, the number of couple types considered in the computer implementation was $15 \times 15 = 225$. Given that r was chosen as 12, the number of age classes for both single females and males was 13 and these classes were denoted by $x = 0, 1, 2, \ldots, 12$ for each of the three genotypes. When either a female or male member of a couple reaches the age of 7, the software was so constructed that couples with a female or male of this age dissolve and their partners were sent back to the single population, where subsequently they may become members of a couple again if their ages were in the interval $x \in [2, 6]$. This strategy was used as a practical device to limit the number of couple types under consideration at any time in an experiment involving many time units of evolution. A biological interpretation of this device is that only females and males in the five prime ages of reproduction were allowed to contribute offspring to the evolving population, which could be a policy followed by some animal breeders.

Throughout all experiments reported in this section, the following parameters were held constant. For all couple types, the probability of couple formation per unit time was chosen as 0.9, and the rate of dissolution for reasons other than death for each couple type per unit time was $\theta_{dis} = 0.1$. The probability of mutation per meiosis of allele A to allele a for both females and males was assinged the value $\mu_{12} = 10^{-5}$, and, similarly, the probability of the mutation $a \rightarrow A$ was assigned the value $\mu_{21} = 10^{-6}$. With regard mortality, the infant survival probabilities for the three genotypes AA, Aa and aa were assigned the values $(0.9.0.9, 0.9)$ for both sexes, see section 12.11 for an interpretation of these probabilities and the relationship the α-parameters in the risk function for early life mortality. Just as in the experiments reported in sections 12.12 and 12.13, the β-parameters in the risk function for early life were assigned the value $2/x_m$ for both sexes. Moreover, the Makeham parameters were assinged the values $(0.001, 0.001, 0, 001)$ for the three genotypes of both sexes. Finally, with respect to the Gompertz risk function for deaths at the mature ages, the mode of the distribution for each genotype and both sexes was assigned the value $(6, 6, 6)$ with corresponding standard deviation $(2, 2, 2)$. Briefly, with respect to the components of natural selection just described, there was no selection in the computer experiments reported in this section. As was the case for the computer experiments reported the foregoing sections of this chapter, the probability a baby was female was assigned the value $p_f = 100/205$ and the probability the baby was male was assigned the value $p_m = 105/205$. Like all branching type processes considered in this book, the two sex, age dependent process with couple formation under consideration was self regulating or density dependent. Furthermore, the expected number of offspring contributed to the population for each of the three genotypes were dependent on age of a female in her reproductive years as described in previous sections of this chapter.

In experiment 12.14.1, a decision was made to study sexual selection so as to compare the performance of the formulation under consideration with those studied in previous sections of this chapter and chapter 11. For this experiment, the matrix of acceptance probabilities for females was chosen as

$$\boldsymbol{A}_f = \begin{pmatrix} 0.1 & 0.1 & 0.9 \\ 0.1 & 0.1 & 0.9 \\ 0.1 & 0.1 & 0.9 \end{pmatrix}, \qquad (12.14.1)$$

and with respect to males it was assumed that all the elements in the matrix of acceptance probabilities were 1, indicating males choose their sexual partners at random. It should also be mentioned that in the formulation under consideration, the choice of mates did not depend on the ages of the partners but only their phenotypes. Observe that implicit in the structure of the matix $\boldsymbol{A}_f$ was the assumption that the allele A was dominant to allele a so that the genotypes AA and Aa expressed the same phenotype. The total fertility of females of each of the three genotypes, the expected number of offspring produced by each females throughout her reproductive years, were assinged the values $(12, 12, 12)$, which resulted in a rapid population increase from the initial population. Lastly, in the initial population is was assumed that the population was homozygous for the allele A and that in the single population each of the 13 age classes contained 10 individuals for both sexes all of genotype AA. In other words, the initial number of individuals genotypes Aa and aa was 0 for each of the 13 age classes for both sexes in the single population. Similarly, for the case of 225 couple types, it was assumed that the 25 couple types corresponding to matings the type $AA \otimes AA$ each had 10 couples in the initial population, and, moreover, it was assumed the number of couples in each of the remaining couple types was 0. Presented in Figure 12.14.1 are the graphs of the total numbers of each of the three genotypes in the single females population for 6,000 time units of evolution based on the embedded deterministic model. It is suggested that a reader may wish to interpret these time units as years for a short-lived diploid animal species with two sexes and with a life span of about 12 years. From the operational point of view, the software performed very efficiently in that the computer time taken to simulate 6,000 time units of evolution was only about two minutes.

As can be seen from the graphs in this figure, the patterns of the trajectories of the three genotypes under sexual selection are similar to those presented in Figure 12.12.1 except that the rise of the genotype aa to predominance in the population occurred much earlier in the projection in experiment 12.14.1 than in experiment 12.12.1, due to the shorter life span of individuals in experiment 12.14.1 of about 12 years in contrast of life span of 100 years for individuals in experiment 12.12.1. The graphs of

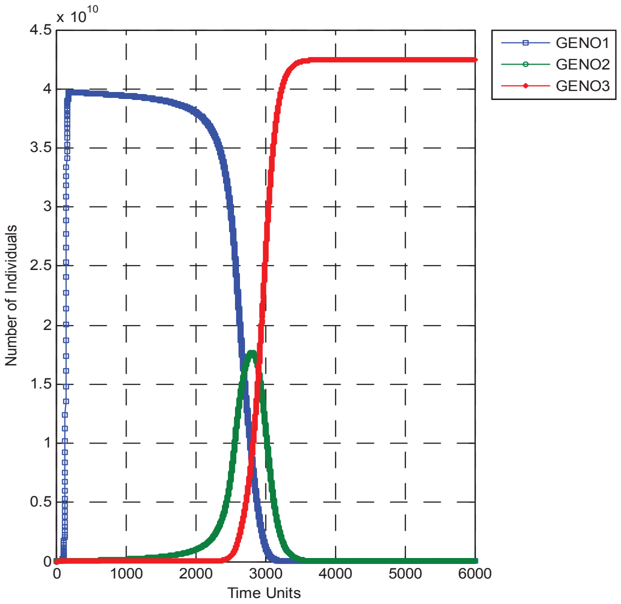

Figure 12.14.1 Graphs of the Trajectories of the Total Number of Each of the Three Genotypes in the Female Population of Singles in Experiment 12.14.1.

the trajectories for the total numbers of each of the three genotypes in the single male population, females in couples and males in couples were very similar to those in Figure 12.14.1 and were, therefore, omitted. This similarity however, did indeed confirm that the genotype aa had risen to predominance in the entire population. Graphs of the age distribution for experiment 12.14.1 were also similar in shape to those presented in sections 12.12 and 12.13 and will, therefore, be omitted for the sake of brevity.

In experiment 12.14.2 there was a departure from the theme that has been followed in many of illustrative experiments reported in this book in that rather than letting a population evolve from a initial population that was homozygous of the genotype AA so that mutant genotypes AA and Aa would arise through mutations, it was assumed that all genotypes were present in the initial population. More precisely, it was assumed that in each of 13 age classes for single females and males each of the three genotypes were represented by 10 individuals. Furthermore, each of the 225 couple types were represented by 10 couples. All the remaining parameter values that were used in this experiment were, however, the same as those in experiment 12.14.1. Thus, the two experiments differed only in the initial conditions. Graphs of the trajectories of the total numbers for each of the three genotypes in the population of single females are presented in Figure 12.14.2 for 6,000 time units of evolution. As can be seen from this figure, the population underwent a period of rapid evolution and, since the genotype *aa* was favored by sexual selection just as in experiment 12.14.1, the genotype *aa* rapidly became predominant in the population within about 700 time units of evolution. Unlike the trajectories for genotypes *AA* and *Aa* in experiment 12.14.1, the total numbers of these genotypes did rise to significantly large numbers in experiment 12.14.2. Like the outcome of experiment 12.14.1, in this experiment genotype *aa* became predominant in the entire population.

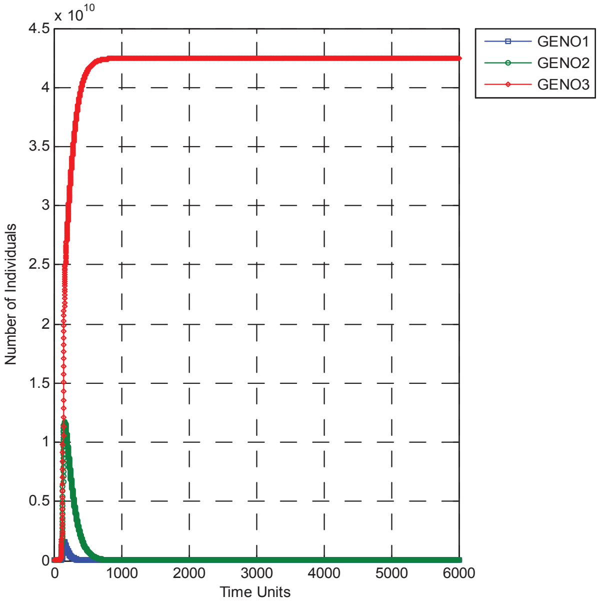

Figure 12.14.2 Graphs of the Trajectories of the Total Number of Each of the Three Genotypes in the Female Population of Singles in Experiment 12.14.2.

Experiment 12.14.3 was designed to test whether a reproductive advantage of genotypes *aa* over the other two genotypes was sufficient for this genotype to become predominant in the population within 6,000 time units of evolution. To carry out this test, the expected total number of offspring contributed to the population by females in their reproductive years were assigned the values $(8, 8, 12)$, respectively, for the three genotypes AA, Aa and aa. It was also assumed that both females and males choose their sexual partners at random with respect to age and genotype so all probabilities in the matrix of acceptance probabilities for both females and males were assigned the number 1. All other parameters of the model had the same values as those for experiment 12.14.1, including the initial condition that all single females and males as well as those who were members of couples in the initial population were of genotype AA. Contained in Figure 12.14.3 are graphs of the trajectories for the total numbers of individuals of each of the three genotypes for 6,000 time units of evolution.

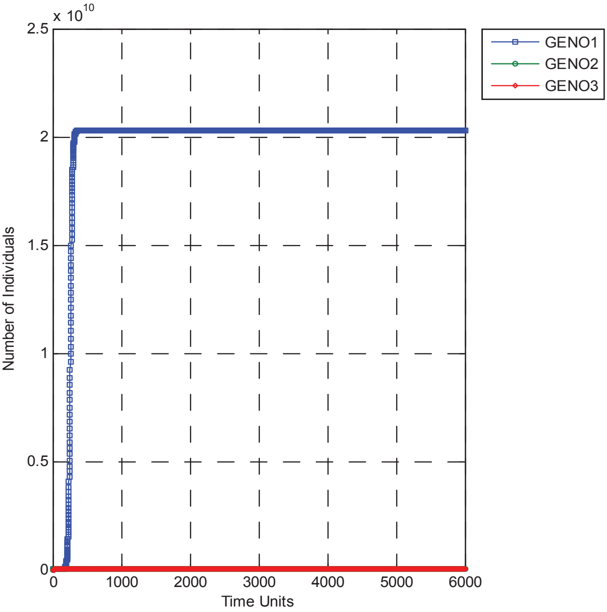

Figure 12.14.3 Graphs of the Trajectories of the Total Number of Each of the Three Genotypes in the Female Population of Singles in Experiment 12.14.3.

As can be seen by an inspection of this figure neither of the mutant genotypes Aa or aa rose to significantly high numbers during the 6,000 time units of evolution considered in the experiment. On the other hand however, the initial genotype AA quickly rose to predominance in the population and converged to a limit within 400 to 500 time units of evolution. Like that for the genotype AA the total number for each of the mutant genotypes also converged to constants. Evidently, because the population was self regulating with respect to total population size, the quick rise to predominance of genotype prevented the genotype aa from reaching sufficiently large numbers to become predominant in the population even though individuals of this genotype had a reproductive advantage.

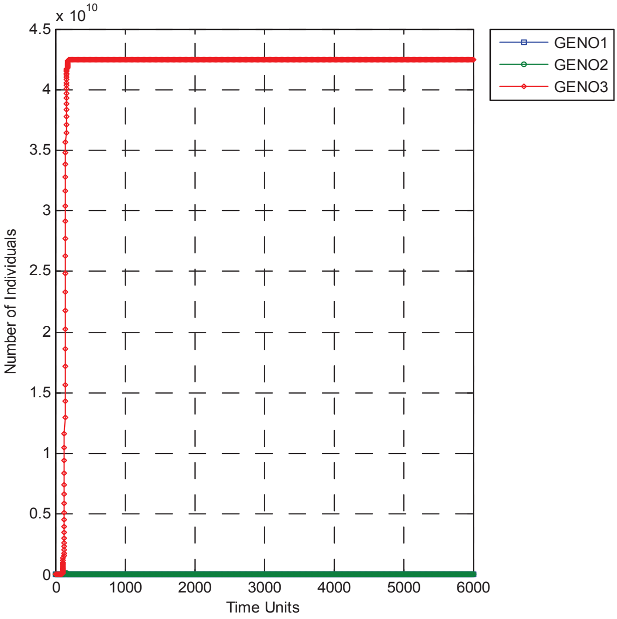

Figure 12.14.4 Graphs of the Trajectories of the Total Number of Each of the Three Genotypes in the Female Population of Singles in Experiment 12.14.4.

In an attempt to find conditions under which the genotype aa with a reproductive advantage would become predominant in the population, in experiment 12.14.4 the same parameter values as those in experiment 12.14.3 were used but the initial conditions were the same as in experiment 12.14.2. That is, in the single initial population of females and males each of the three genotype were represented by 10 individuals for each of the 13 age classes. Similarly, each of the 225 couple types in the initial population contained 10 couples.

Presented in Figure 12.14.4 are graphs of the total numbers of each of the three genotypes in the single female population for 6,000 time units of evolution. Unlike the graphs displayed in Figure 12.14.3, in Figure 12.14.4 the genotype *aa* with a reproductive advantage quickly rises to predominance and converges to a constant within 600 time units of evolution. Throughout the projection, however, the of individuals of genotypes *AA* and *Aa* remained at relatively low levels. This experiment suggested that, given a level initial playing field, a genotype with a reproductive advantage will eventually rise to predominance, even if the population is self regulating in population size. It is also of interest to note that a stable limit point attained by the non-linear difference equations making up the embedded deterministic model seem to depend on the initial conditions.

For those readers who are interested in looking at historical mortality data and demographic structures of human populations, it is suggested that the works Alderson (1981), Keyfitz and Flieger and Rogot *et al.* (1988) be consulted.

Bibliography

[1] Alderson, M. (1981) **International Mortality Statistics**. Facts on File, Inc. New York.

[2] Barbu, S. and Limnois, N. (2009) **Semi-Markov Chains and Hidden Semi-Markov Models Towards Applications - Their Use in Reliability and DNA Analysis**. Lecture Notes in Statistics, Springer.

[3] Charlesworth, B. (1980) **Evolution in Age Structured Populations**. Cambridge University Press, Cambridge, London and New York.

[4] Keyfitz, N. and Flieger, W. (1968) **World Population**. The University of Chicago Press, Chicago and London.

[5] Mode, C. J. (1985) **Stochastic Processes in Demography and Their Computer Implementation**. Springer, Berlin, Heidelberg, New York, and Tokyo.

[6] Mode, C. J. and Sleeman, C. K. (2000) **Stochastic Processes in Epidemiology, HIV/AIDS, Other Infectious Diseases and Computers**. World Scientific, Singapore, New Jersey, London, Hong Kong.

[7] Rogot, E. *et al.* (1988) **A Mortality Study of One Million Persons by Demographic, Social and Economic Factors: 1979-1981 Follow Up**. U.S. Public Health Service, National Institutes of Health, NIH Publication No. 88–2896.

Chapter 13

An Overview of the History of the Concept of a Gene and Selected Topics in Molecular Genetics

13.1 Introduction

In the foregoing chapters of this book, genes and mutations at some locus were defined on an abstract Mendelian level such that mutations among alternative forms of genes, alleles, could occur with given probabilities per generation. Within this framework, the components of natural selection were supposed to act in such a way that the most advantageous allele, which arose as a mutant, would in time become predominant in a population as it evolved over generations or, in some cases, it may be maintained in a population through a what is known as balancing selection. In passing it was also tacitly recognized that genes at different loci could also be the structures upon which natural selection acts, but, because of the complexities that arise in processing multidimensional arrays when statistically summarizing the results of a Monte Carlo simulation experiment, the implications of such formulations were not explored. Even though computer simulation experiments on the quantification of mutation and selection conducted within such frameworks were helpful in illuminating ways in which the components of natural selection may to drive the changing genetics of a population, they provide little or no insights into what is happening at the molecular level.

Accordingly, the purpose of this chapter is to provide an overview of recent and past developments in molecular genetics with a view towards providing frameworks to help clarify our thinking and experimentation about the mechanisms upon mutation and natural selection may act at the molecular level. This overview begins with a review of the concept of a gene and then proceeds to a view of how this concept has changed in light of micro array experiments with samples of DNA. Then, as the concept of a gene progresses, selected topics from molecular genetics are introduced to provide a more firm basis for our thinking about the driving forces of evolution, mutation and selection. Finally, three specific examples of genes at the molecular level are given in an attempt to anchor our thinking in some actual observations that have been made by experimenters working in the field of molecular genetics. To readers of books on mathematical genetics, in which little attention has in the past be given to molecular genetics, it should be made clear that the material presented in this chapter was motivated by a desire and an attempt to develop a more complete framework in which to further the development of mathematical genetics. To specialist in molecular genetics, however, the overview attempted in this chapter will, no doubt, be lacking in essential details and for such readers, who may glance at the contents of this chapter, we request their forbearance.

13.2 A Brief History of the Definition of a Gene

The history of the definition of a gene began with the publication of Mendel's experiments in 1865 with peas. An English translation of the original paper, which was written in German, may be found in the appendix of the well-known textbook Sinnott, Dunn and Dobzhansky (1950). Mendel's seminal work, however, remained unnoticed until about 1900, when De Vries in Holland, Correns in Germany and von Tschermak in Austria found Mendel's forgotten paper and recognized its importance. An account of this story may be found in chapter 2 of the book by Sinnott *et al.* cited above. The main conceptual thrust of Mendel's work was the idea that a unit of inheritance, which is now spoken of abstractly as a gene, was some particulate entity that was passed was on from generation to generation from parents to their offspring. Indeed, with the exception of chapters 6, 7 and 8, in which the mutational process of nucleotide substitution and its ramifications were considered, the notion of the particulate nature of the gene has been tacitly used in the formulation of the various stochastic process discussed as well as in the computer experiments devoted to the quantification of mutation and selection, two widely recognized processes underlying Darwinian evolution.

The next influential period on the development of the concept of the gene started in about 1910 with the study of Mendelian segregation and recombination of mutations in experiments with *Drosophila melanogaster* by T. H. Morgan and his students, which were reported in the influential book Morgan *et al.* (1915). In the view of the gene developed by this group, genes were thought of as beads on a string such that a bead was the site or locus of a gene and the distances among genes on the string was expressed in terms of centimorgans as mentioned in the treatment of genetic recombination in chapter 2. Actually, the loci making up an observed linkage group, were thought of being located on some chromosome. However, it was not until 1929, when Barbara McClintock (1929) with her cytogenetic studies of maize, showed genetic linkage corresponded to physical locations on chromosomes.

Another influential definition of the gene rose to prominence in the 1940s as a result of the work of Beadle and Tatum (1941). These authors, who worked with the metabolism of the fungus, Neurospora, showed that mutations could effect biochemical pathways adversely. It was thought that each step in a biochemical pathway was catalyzed an enzyme which was known to be a protein. This led to the idea of one gene one enzyme view, which subsequently become known as one gene one polypeptide. In this view, each step in a biochemical pathway was catalyzed by an enzyme and a gene contained the information to code for the particular protein in each step. As the years went by this view was made more explicit and mechanistic.

During the 1950s a view developed that a gene was a physical molecule. Experiments that gave credence to this view were those in which H. Muller (1927) demonstrated that heredity has a physical and molecular basis by mutating genes with x-ray radiation. Other evidence that supported this view was the work of Griffith (1928), who demonstrated that something in the virulent but dead Pneumococcus strain could be taken up by live non-virulent strains and transform them into virulent bacteria. In 1955, Hershey and Chase showed that the substance referred to by Griffith was actually DNA and not protein. Furthermore, the notion that the product of a gene was a diffusible substance was used to define genes in the early years, using ideas from genetics in applications to bacteriology. In these experiments an operational view of a gene was that it was a cistron, a region of DNA defined by mutations that in trans configuration could not genetically complement each other. For further details consult Benzer (1955).

Following the publishing of the Watson and Crick (1953) paper, in which they proposed a double helix model for the structure of DNA, in the 1960s the definition of a gene evolved into the idea of transcribed code. The double helix model of DNA with two strands of paring bases provided a basis as to how one strand could serve a template for the copying of another strand and how mutations may occur, the substitutions of one nucleotide for another, as errors in the copying process. During the 1960s and until the present day, molecular biology developed at a rapid pace, and one of the outcomes of this rapid development was the demonstration that RNA transcripts, consisting strings of three letter codons, coded for a sequence of amino acids making up a protein. For further details, the papers of Nirenberg *et al.* (1965) and Soll *et al.* (1965) may be consulted. In subsequent sections of this chapter more details will be provided on the process of transcription. In 1958 Crick (1958) had outlined the flow of information from a nucleic acid to a protein, which began the development of the central dogma of biology. The central dogma will be discussed in more detail in subsequent section. However, even in this early development stage of the central dogma, it was known that some products resulting from the DNA transcription process were ribosomal and transfer RNA. Moreover, in RNA viruses it was known that genes were composed of RNA. In summary, during 1960s, the prevailing view of a gene was that it was a code residing on a nucleic acid that gave rise to a functional product.

With the development of cloning and sequencing techniques in 1970s, combined with a knowledge of the genetic code, another evolutionary phase of molecular biology consisted of numerous observations that provided extensive information on how genes were organized and expressed. This led to the idea of viewing a gene as an open reading frame (ORF). The first gene to be sequenced was that from the bacterial phage MS2 and it was also the first organism to have its genome completely sequenced, see Fiers *et al.* (1971) and (1976). The parallel development of computational tools led to algorithms for the identification of genes based on their sequence characteristics, or the arrangement of the nucleotides. For a review of this development see Rogic *et al.* (2001). This development created a new concept of a gene, which was defined in terms of its predictive sequence rather than a genetic locus responsible for its phenotype, see Griffiths and Stotz (2006). In many cases, a DNA sequence could be used to infer structure and function for a gene and its products. As a result of this development, the identification of most genes in sequenced genomes was based on either their similarity to other known genes, or on a statistically significant signature of a protein-coding sequence. Quite often, a gene was identified as an annotated ORF in a genome under study, see Doolittle (1986).

During the 1990s to 2000s, there was another shift in the definition of a gene as an annotated genomic entity enumerated in data banks. Currently, the definition of gene utilized by scientific organizations that annotate genomes still relies on the sequence view. For example, a gene was defined by the Human Genome Nomenclature Organization as a "*DNA* segment that contributes to phenotype/function". In the absence of demonstrated function, a gene may be characterized by sequence, "transcription or homology", see Wain *et al.* (2002). There are also other descriptions of a gene, for example, the Sequence Ontology Consortium called a gene a "locatable region of a genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions and/or other functional sequence regions", see Pearson (2006).

With the first sequencing of the Haemophilus influenza genome and then the human genome, Fleischmann *et al.* (1995), Lander *et al.* (2001) and Venter *et al.* (2001), a very large amount of sequenced *DNA* became available for which the above definitions could be applied. There was, in fact, much popular interest in counting the number of genes in various organisms. Even at that time, however, it was pointed out that the enumerated genes over emphasized traditional protein coding genes, and when the number of genes in the human genome was reported in 2003 as a climax of the Gene Sweepstakes, it was acknowledged that too little was known about *RNA* coding genes. In this connection the ensemble view of the gene was emphasized specifically in the rules of the Gene Sweepstakes as "alternatively spliced transcripts all belong to the same gene, even if the proteins produced are different." In a subsequent section, the phenomenon of alternatively spliced transcripts will be treated in more detail.

There is also a metaphor in current usage, especially among those in the bioinformatics community, that views genes as subroutines in the genomic operating system. Evidently, this metaphor arose out of the fact that the counting of genes in a genome was and is a large-sale computational endeavor and that genes fundamentally deal with the processing of information that is coded in the *DNA*. This has led some people in the computational biology community to use the description of a formal language that is also used to describe computer programs with precise syntax such as upstream regulation, exons and introns, see Searls (1997), (2001) and (2002). The metaphor of thinking of the genome as the operating system for a living organism seems be useful in thinking about the concept of a gene in the sense that nucleotides of a genome are put together into a code that is executed through the process of transcription and translation and that genes are subroutines that are called in the processes of transcription. In a subsequent section, more details will be given regarding the terms exon, intron, transcription and translation.

13.3 Overview of Transcription and Translation Processes

Before developing a more up to date concept of a gene, it will be helpful to deepen the reader's awareness of some basic knowledge concerning the processes of transcription and translation. Accordingly, the purpose of this section is to provide of an overview of the material on the web site along with illustrative examples. If this web site is not available to a reader, other web sites may be consulted as well as books on genetics and biochemistry. In this connection the reference, Snustad and Simmons (2006), is a good source for material on DNA and RNA. The process of gene expression occurs in two steps:

- **Transcription** $=$ $DNA \to RNA$
- **Translation** $=$ $RNA \to$ **Protein**

These two steps together make up the "central dogma" of biology: $DNA \to RNA \to$ **Protein**.

To further elaborate the central dogma, the discussion will begin with a symbolic description of the process of transcription. Below is an illustrative overview of a section of a double stranded DNA molecule.

$5'$	·	·	·	A	T	G	G	C	C	T	G	G	·	·	·	$3'$
$3'$	·	·	·	T	A	C	C	G	G	A	C	C	·	·	·	$5'$

The upper stand in the $5' \to 3'$ direction is referred to by some authors as the sense strand of DNA. The lower strand in the $3' \to 5'$ direction is known as the antisense strand. The numbers 3 and 5 refer to the conventional numbering of five carbon atoms in the five carbon sugars called pentoses that are part of a DNA molecule.

The transcription process begins with the antisense strand of the double stranded DNA molecule and is illustrated symbolically below

$3'$	·	·	·	T	A	C	C	G	G	A	C	C	·	·	·	$5'$

$$\Downarrow$$

$5'$	·	·	·	A	U	G	G	C	C	U	G	G	·	·	·	$3'$

where the symbol $\Downarrow$ denotes transcription. As a means of symbolically distinguishing DNA from RNA, RNA will be denoted by double bars. The strand in the $5^{'} \to 3^{'}$ direction represents a molecule of RNA transcribed from the antisense strand of a DNA molecule in the $3^{'} \to 5^{'}$ direction. Note, in this RNA molecule, the base U called uracil has replaced the base T called thymine in the DNA sense strand in the above section of a DNA molecule. In general, whenever a molecule of RNA is transcribed from a strand of DNA, the base T is replaced by the base U.

The next step in the central dogma is the process of translation in which a messenger RNA molecule, symbolized by $mRNA$, codes for amino acids, the building blocks of proteins, by utilizing a three letter code consisting of bases. In this illustrative example, a section of a RNA molecule is illustrated by 9 bases. For example, the first three bases AUG code for the amino acid Methionine symbolized by $AUG \Longrightarrow Met$. The next three bases GCC code for the amino acid Alanine symbolized by $GCC \Longrightarrow Ala$. Finally, in this example the last codon is UGG, which codes for the amino acid Tryptophan symbolized by $UUG \Longrightarrow Trp$. In this illustrative example, the translation process produces a peptide consisting of three amino acids symbolized by

$$Met - Ala - Trp.$$

Proteins are three dimensional structures consisting of peptide chains which fold in numerous ways, and have physical properties which, among other things, depend on the ways a protein is folds. More detailed information on translation may be found on the web site as well in text books on genetics and biochemistry. In particular, a table is presented which identifies the amino acid coded for by each of the 64 possible combination of three letter codes, using the symbols A, U, G and C. Some of these codes are stop codons, which halt the process of transcription. It should also be mentioned that transcription may occur, using the sense strand of a DNA molecules as a template for a RNA molecule. A symbolic representation of the process is very similar to that demonstrated above for the antisense strand of a DNA molecule and will, therefore, be left as an exercise for the reader.

Interestingly, proteins also play a role in the process of transcription, whereby a single strand of DNA serves as a template for the formation of a strand of RNA as illustrated above. In this connection some 50 different protein transcription factors, which bind to promoter sites usually on $5^{'}$ side of a gene to be transcribed, have been described. An enzyme, called RNA polymerase, binds to this complex of transcription factors and working together they open the double DNA helix to begin the transcription process. From the chemical point of view, this transcription process is rather complicated so no further details will be given here. An interested reader may, however, consult the web page cited above as well as text books on genetics and biochemistry

In recent years, however, the central dogma concerning the nature of the gene in the sense that it codes only for a protein has come into question as a result of the ENCODE and other projects in which the functional aspects of DNA are being investigated following the sequencing of the human and other genomes. Evidently, problems arise in that molecules of RNA, which arise through a transcription process, play important roles in many cellular processes both in the nucleus and the cytoplasm. The stating point of the discussion is that DNA serves as the template for the synthesis of RNA just as it does for its own replication.

Before proceeding with the discussion, it will be helpful to discuss several types of RNA that are synthesized in eukaryotic cells and, evidently, are products of the process of transcription. One of these products is messenger RNA, which is denoted by $mRNA$ and is later translated into a polypeptide. Another product is ribosomal RNA, which is denoted by $rRNA$, and will be used in building ribosomes, which are part of the machinery for synthesizing proteins by translating $mRNA$. There is also transfer $RNAs$, denoted by $tRNA$, which carry amino acids to a growing peptide chain. Another form of RNA is called small nuclear RNA, which is denoted by $snRNA$. Segments of DNA, which transcribe for $mRNA, rRNA$ and $tRNA$, produce large precursor molecules, called primary transcripts, that must be processed in the nucleus to produce functional molecules for export to the cytoplasm. Some of these processing steps are mediated by $snRNAs$. There are also small nucleolar RNA, symbolized by $snoRNA$, which are $RNAs$ in the nucleolus and have several functions, see the web page cited above. There are also micro $RNAs$, which consists of a small number of nucleotides, i.e., 22 or more, which appear to regulate the expression of $mRNA$ molecules. Such micro $RNAs$ are denoted by $miRNA$. Finally, there is a RNA denoted by $XIST\ RNA$, which inactivates one of the two X chromosomes in female vertebrates. Again the web page cited above may be consulted for more details.

The only molecules that encode polypeptides are messenger $RNAs$. The number of nucleotides in these molecules depends on the number and kind of amino acids that make up a polypeptide. All other classes of RNA molecules, some of which are not mentioned in this discussion, are called non-coding $RNAs$ and symbolized by $ncRNAs$. It has been estimated that 2/3 of the transcription products in the nucleus are $ncRNAs$. At the present time, not much is known about the function or functions of *ncRNAs* but they are increasingly a subject of research. Interestingly, in the paper by Ponjavic *et al.* (2007) statistical evidence was presented for selection within long non-coding *RNAs*.

That *RNA* plays very basic roles in biology has been presented convincingly by T. R. Cech, a 1989 Nobel Laureate in Chemistry, see Cech (2004), who states that within the last few years, *RNA* research has reached new heights, and it has become clear that *RNA* catalysis plays a much more central role in biology than previously thought. Moreover, *RNA* often controls the expression of genes, which is another role that has been thought to be mostly the role of proteins called repressors and transcription factors. There are also *RNA* machines in cells called ribosomes that translates the message encoded in *mRNA* to proteins, which do most of the work in biology. Proteins can be hormones such as insulin or sex hormones, while others may catalyze the digestion of food. Each messenger *RNA*, *mRNA*, codes for a specific protein, and the ribosome is remarkable in that it can decode a very large number of different *mRNAs*.

The ribosome is also a remarkable catalyst composed of three *RNA* molecules, or four in some species, as well as dozens of proteins. Over the last three decades, a growing body of evidence has been developed that *RNA* may lie at the catalytic heart or active site of the ribosome, with proteins playing supporting roles. In 2000, an atomic-level picture of ribosome emerged, showing that the complex fold of the *RNA* molecules buttressed by numerous proteins. One structure showed the site where amino acids are strung together into proteins. In technical terms, this site is called the peptidyltransferase center. Furthermore, the atomic-level image of this site showed that it was composed of *RNA*, with no proteins in the vicinity. Such images provide the most direct evidence that the ribosome is indeed a *RNA* catalyst. Such catalysts are also known as ribozymes. It is suggested that a reader also consult the paper Cech (2004) as well as other sources for more details on the roles *RNAs* play in various cellular processes.

This brief and over simplified overview of the literature on the processes of transcription and translation strongly suggest that the central dogma of biology needs to be extended to include the roles of transcribed *RNA* plays in cellular processes.

13.4 Pre-processing Messenger RNA-Introns and Exons

More background information will be needed to develop a comprehensive concept of a gene. Like the material presented on the processes of transcription and translation, most the material presented in this section has been abstracted and revised from the web site. Most transcripts produced in a nucleus of a cell must undergo several processing steps to produce functional RNA for export to the cytoplasm. Evidently, many examples of this are known, but in this section attention will be confined to a view of steps that occur in the processing of pre-messenger RNA, $pre - mRNA$, to $mRNA$. Most regions of DNA in eukaryotic cells, which contain three letter codes for proteins, are partitioned into segments. Those segments of DNA that are transcribed but not translated into proteins are called introns. Those segments of DNA that contain codons for amino acids in a protein are called exons. An example of such a region of DNA in humans is that for dystrophin, which has 79 exons. Boys with muscular dystrophy carry a mutant form of this gene. It has also been reported that regions of DNA that code for $rRNA$ and $tRNA$ also contain introns. In general, introns tend to be much longer than exons. For example, the average exon in eukaryotes has been reported to be 140 nucleotides in length. In general, however, introns may be rather long, and it has been reported that the length of one human intron is 480,000 nucleotides.

The removal of introns and the splicing of exons together are among the steps in synthesizing $mRNA$. Consider, for example, the symbolic region of DNA

Exon 1	Intron A	Exon 2	Intron B	Exon 3

consisting of three exons and two introns. In this representation each exon and intron may contain different numbers of the bases A, T, G and C. The first step in the pre-processing of RNA is the of transcription of a strand of DNA into a strand of RNA, which is symbolized below:

$5^{'}$	Exon 1	Intron A	Exon 2	Intron B	Exon 3	$3^{'}$

In this representation of a strand of *pre-mRNA*, the cells in the table have been outlined with double bars to indicate that the process of transcription has occurred.

The next step in the pre-processing of RNA is called capping. In this process a modified guanine, G, is attached to the $5^{'}$ end of the *pre-mRNA* strand as it emerges from *RNA* polymerase II, which is a enzyme (protein), symbolized by *RNAP II*. It is thought that the cap protects *RNA* from being degraded by enzymes that degrade *RNA* starting with the $5'$ end. It is also thought that the cap acts as an assembly point for proteins needed to recruit a small unit of the ribosome, which is necessary to begin the process of translation. The capping process is denoted symbolically below:

G	$5'$	Exon 1	Intron A	Exon 2	Intron B	Exon 3	$3'$

The next step in the pre-processing of *mRNA* is the step-by-step removal of introns and splicing of the remaining exons. This step occurs as the *pre-mRNA* continues to emerge from *RNAP-II*. The excision of introns and the splicing of exons are symbolized below:

G	Exon 1	Exon 2	Exon 3

The next step in the pre-processing of *mRNA* is called Polydenylation, which entails the synthesis of the poly(A) tail. The tail consists of stretch of adenine *A* nucleotides. As a special poly(*A*) attachments site in the *pre-mRNA* emerges from *RNAP-II*, the transcript is cut there, and the poly(*A*) tail is attached to the $3'$ end. This step completes the pre-processing of *RNA* into *mRNA*, which is now ready for export to the cytoplasm. At this juncture, the remaining transcript is degraded and the *RNA* polymerase leaves the *DNA*. The process of Polydenylation, which is the last step in the pre-processing of *mRNA*, is symbolized below:

$5'$	G	Exon 1	Exon 2	Exon 3	Poly(*A*)	$3'$

The process of cutting and splicing must be done with great precision. For, if even one nucleotide is left over from an intron or one is removed from an exon, the reading frame from that point of will be shifted, which results in the production of new codons that specify a totally different sequence of amino acids from that point to the end of molecule, which may often end prematurely when the shifted reading frame generates a *STOP* codon. The removal of introns and splicing of exons are done by complexes called spliceosomes, which are composed of 5 *snRNA* molecules and about 145 different proteins. The introns in most *pre-mRNAs* begin with *GU* and end with *AG*. Evidently, the stretches of two nucleotides assist in guiding the spliceosome.

There is also a phenomenon called alternative splicing, which occurs in the processing of *pre-mRNA* for many proteins and proceeds along various paths or under different conditions. For example, in the early differentiation of a B cell, a lymphocyte that synthesizes an antibody, the cell first uses an exon that encodes a transmembrane domain that causes the molecule to be retained at the cell surface. Later on, the B cell switches to using a different exon, whose domain enables the protein to be secreted from the cell as a circulating antibody molecule.

Apparently, the phenomenon of alternative splicing also provides a mechanism for producing a wide variety of proteins from a relatively few regions of DNA that contain exons. Usually, these regions are called genes. The question as to how many genes there are in the human genome remains open, but some estimates suggest that there is about 23,000 genes, which probably codes for ten times that number of proteins. It has been estimated that 92-94% of our genes produce *pre-mRNA* that are alternatively spliced. There is also evidence that patterns of alternative splicing differs consistently among tissues, which suggests that this process is regulated. However, whether all the products of alternative splicing are functional or that many are simply the outcome of an error-prone stochastic process remains to be seen.

The acronym UTR stands for untranslated region. There is evidence that alternative splicing not only provides different proteins from a single gene but also different $3^{'}$ $UTRs$ and $5^{'}$ $UTRs$. Even though these regions are not translated, they contain signals, for example, that determine where in the cell a protein will accumulate. An example of this phenomenon is the $3^{'}$ UTR of the bicoid gene in Drosophila that directs the $mRNA$ to the anterior of the embryo.

One of the most dramatic examples of alternative splicing is the $DESCAM$ gene in Drosophila. This gene contains 116 exons of which only 17 are retained in the final $mRNA$. Some exons are always included in $mRNA$ that is translated into protein; while in other cases they are selected from the array of 116 exons. It has been estimated that this system is able to produce 38,016 different proteins, and, as it turns out, over 18,000 different proteins have been observed in Drosophila hemolymph. The $DESCAM$ proteins appear to be involved in establishing a unique identity for neurons, guiding neurons to their proper destination in embryonic stage and probably in the recognition of and phagocytosis of invading bacteria.

Humans also have a $DESCAM$ gene located on chromosome 21. When this gene is present in three copies, it seems likely that it is responsible for some symptoms of Down's syndrome. It also accounts for the full name of this disorder, Down's Syndrome Cell Adhesion Molecule. It has also been suggested that the incredible diversity of synaptic junctions in mammalian central nervous systems, which have been estimated to be of the order 10^{14}, is mediated by alternative splicing of limited number of gene transcripts.

It appears, therefore, that whether a particular segment of *RNA* will be retained in an exon or excised as an intron can vary under different circumstances such as what type of cell the gene is in, what stage of differentiation that a cell is passing through and on what extracellular signals that the cell is receiving. Observations of this kind point to the conclusion that alternate splicing pathways must be closely regulated by such complexes as transcription factors and other complex regulatory structures that bind to *DNA*, which are sometimes called motifs.

A question that naturally arises is what the evolutionary significance of split genes is? One view as to their origin is that eukaryotic genes have been assembled from smaller primitive genes, which are the exons observed in living species. Some proteins, such as the antibodies mentioned previously, are organized into a set of separate sections or domains, with a special function to perform in the complete module whose parts are encoded by separate exons. This presence of separate functional parts in an antibody molecule encoded by different exons makes it possible to use these units in different combinations. Thus, it appears that a major evolutionary benefit of spit genes, *i.e.*, regions of *DNA* containing codons for proteins, is that the simply provide opportunities for making many different proteins from a single coding region through the mechanism of alternative splicing.

Moreover, some of these combinations may have evolutionary advantages to the individuals that carry them in terms of such components of natural selection as selection of mates, reproductive success and competitive ability for scarce resources. In this connection, some specific examples, connecting the functioning of such combinations at the molecular level with some component or components of natural selection, would be of great interest.

In this and the preceding sections of this chapter, only a brief overview of transcription, translation, introns and exons have been given. For a more detailed account of a discussion of phenomena that complicate the concept of a gene, it is recommended the a reader consult the review paper Gerstein *et al.* (2007). In particular, a perusal of the material and references in table 1 of this paper can be very informative. The evidence outlined in this section also points to the need to examine and revise the central dogma of biology.

13.5 Difficulties with Current Concepts of a Gene

A working paradigm for the mechanism underlying the regulation of a gene was provided by Jacob and Monod (1961) in their study of the lac operon of *Escherichia coli.* This mechanism consisted of three parts: a region of DNA made up of sequences coding for one or more proteins, a promoter sequence for the binding of RNA polymerase and an operator sequence to which regulatory genes bind. Subsequently, other sequences were found that could affect almost every aspect of gene regulation from transcription to $mRNA$ degradation and post-translational modification. It was observed that such regulatory sequences could occur within coding sequences as well as in flanking regions.

Moreover, for the case of enhancers and related structures, such segments of DNA could be far from the coding sequence in terms of counts of the number of nucleotide separating them. These regulatory elements far from the sequences that contain codes for proteins made the concept of gene as a segment of DNA with a contiguous and compact set of bases problematic.Currently, regulation is an integral part of definitions of a gene. One current text book on molecular biology defines a gene as the entire set DNA segments of a nucleic acid sequence that are necessary for the synthesis of a functional polypeptide or molecule of RNA, see for example, Lodish *et al.* (2000).

Evidently, this definition implies that the DNA sequences in a gene would include not only those coding for $pre\text{-}mRNAs$ and its flanking control regions but also for enhancers that are distant along a DNA sequence. However, even though they are distant in terms of base pairs, they may be close in terms of Euclidean distance in the three-dimensional chromatin space defined within a framework of Cartesian coordinates.

As the sequencing of genes, $mRNA$ and whole genomes progressed, it became apparent that the simple operon model proposed Jacob and Monod turned out to be applicable to only prokaryotes and their phages, the viruses that attack them. It was found that Eukaryotes were different in many respects with respect to genetic organization and information flow. For example, the model of genes as non-overlapping and continuous sequences of nucleotides was shown to be incorrect by the precise mapping of coding sequences of genes.

It was, in fact, found that some genes overlap and share the same DNA sequence in different reading frames or even on the opposite strand. This discontinuous structure of genes allows for the possibility that one gene may be completely contained within another's intron, or one gene may overlap with another on the same strand without sharing any exons or regulatory sequences.

The phenomenon of gene splicing was discussed in section 13.4 in connection with the preprocessing of $mRNA$. This phenomenon was discovered in 1977 and was reported in the papers Berget *et al.* (1977), Chow *et al.* (1977) and Gelinas *et al.* (1977). As discussed in section 13.4, as a result of alternative splicing, one genetic locus could code for many $mRNA$ transcripts according to the type of cell and stage of development of an individual. This and other discoveries further complicated the concept of the gene.

For example, as an operational definition of a gene, Celera, the private company that sequenced the human genome, defined a gene as "a locus of co-transcribed exons", see Venter *et al.* (2001). Another definition of a gene was used in Ensembl's Gene Sweepstakes web page, where a gene was defined as "a set of connected transcripts". Evidently, in this definition the word "connected" meant sharing at least one exon. Implicit in this definition is the idea that a set of transcripts may share a set of exons, but in may happen that a given exon is not shared by all transcripts in a set.

There is another phenomenon called trans-splicing in which two separate $mRNA$ molecules are spliced by a process termed ligation, which further complicates our understanding of the gene, see Bumenthal (2005). Examples of this phenomenon include transcripts from the same gene, a gene on the opposite DNA strand and even on a separate chromosome. Such examples make it clear that the classic concept of a gene as a locus no longer applies for gene products whose DNA sequences coding for these products are scattered across the genome. Furthermore, there are recent studies that have reported on a phenomenon called tandem chimerism, where two consecutive genes are transcribed into a single RNA molecule, see Akiva *et al.* (2006) and Parra *et al.* (2006). The further processing of such RNA molecules into $mRNA$ and a subsequent translation can lead to a new fused protein composed of parts of two ancestral proteins.

The concept of a gene as the "ORF sequence", as was practised in the 1980s and onward, made it clear that there were large regions on non-genic DNA in the genomes of eukaryotes and especially in the human genome. Without a knowledge as to how these regions functioned, the apparent lack of function led some to label these regions as "junk DNA", see Ohno (1972). Further evidence for this view was obtained when the human genome was sequenced, where it was shown that only about 1.2% of the bases making up the DNA code for exons, see Lander *et al.* (2001) and Venter *et al.* (2001).

As a follow up to these observations, however, it was shown in some pilot functional genomic experiments with human chromosome 21 and 22 that an appreciable amounts of putative "junk DNA" was transcribed, see Kapranov *et al.* (2002) and Rinn *et al.* (2003). Furthermore, in comparisons of the human, dog, mouse and other vertebrate species genomes, it was shown that a large fraction of so called "junk DNA" was conserved in the sense that these species had large regions of DNA in common in their genomes. Evidently, since the time these species had diverged natural selection has conserved substantial sections of their genomes, see Waterston *et al.* (2002) and Lindblad-Toh *et al.* (2005) for more details. Such observations suggest that regulatory sequences of DNA may lie within the "junk DNA" which was swept along in these species by natural selection as they evolved.

13.6 Acronyms Used in Applications of Tiling Array Technology

The acronym ENCODE stands for ENCyclopedia Of DNA Elements, see, which is a project supported by the U. S. National Institutes of Health and involves a consortium of research groups working in molecular biology, genetics, computer science, biostatistics and other disciplines at various universities and other organizations. In what follows, the word genome refers to the human genome unless stated otherwise. Among the technologies used by these groups is that of tiling arrays. Tiling array technology has been developed from microarray technology during the past 5 or so years and has been described in the paper by Lang *et al.* (2008). Further information on this technology may be obtained by entering the phrase "tiling array technology" into a search engine for the world wide web. The literature contains many acronyms and other abbreviations, which are often not defined and make it difficult and laborious to read for first time readers who wish to explore and comprehend this literature. Therefore, as an aide to assist readers in comprehending the material in this and other sections, several widely used acronyms will be defined and discussed.

Among these acronyms is the symbol $cDNA$, which stands for complementary DNA or copy DNA and can be either single or double stranded. This form of DNA is synthesized in vitro from a $mRNA$ template using reverse transcriptase, an enzyme found in nature. The synthesis of *cDNA* is just one among many cases in applications of biotechnology in which enzymes and other chemical compounds present in nature are used to design and expedite experiments. The *cDNA* resulting from this process is single stranded, but it is possible to use such single strands as template to synthesize double stranded *cDNA* that is needed in some experiments. The purpose of converting *mRNA* to *cDNA* is that *DNA* is more stable than *RNA* and thus more amenable to experimentation. When one or a few molecules of *cDNA* are available, they may be multiplied by a process referred to a *RT-PCR*, where *RT* stands for reverse transcription and *PCR* stands of polymerase chain reaction. In a word, *PCR* is a laboratory procedure, based on enzymes that occur in nature, that is used to produce many copies of single or double stranded *DNA*, and such *cDNA* can be used to design probes for expression analysis and cloning of the *mRNA* sequences under consideration. The use of full-length *cDNA* technology and other cloning methods played central roles in finding and characterizing many known and top-quality protein coding genes in the human and other genomes. However, these methods work best for finding highly expressed genes, but are less effective in detecting and characterizing other transcription products, which motivated the development of tiling arrays.

Another acronym that is found in the literature is *ChIP-on-chip*, which stands for chromatin immunoprecipitation-on-chip, and is also known as Location Analysis, LA. This technology provides for a high throughput and genome wide identification and analysis of *DNA* fragments that are bound to specific proteins such as histones, transcriptional factors. This technology involves the following procedures: (1) crossing linking *DNA* with proteins, (2) sonication of *DNA* into small pieces, (3) immunoprecipitation of *DNA*-bound proteins with an antibody, (4) purification of *DNA* and (5) hybridization of *DNA* with microarrays. Briefly, the last procedure involves, among other things, the labeling of *DNA* of some known origin with a radioactive or other substance and checking whether it binds to some strand of *DNA* whose origin is unknown. There is another related acronym, *ChIP-PET*, where *PET* stands for paired end di-tag. Moreover, there are software packages for processing and managing *PET* sequence data listed on the world wide web. There is still another acronym that will be used subsequently and is called *RACE* and stands for the rapid amplification of *cDNA* ends as well as the technology used in the various applications of such amplifications of *cDNA*.

Briefly, a *DNA* microarray is built like a computer chip and this small but powerful device allows researches to observe genes in action, which was thought to be impossible only a decade or so ago. Generally speaking, tiling arrays differ from microarrays in two ways. (1) A traditional microarray provides a technology for test procedures that utilize only part of particular chromosome, but applications of tiling arrays are not restricted to relatively small samples of DNA. (2) The possibilities for large-scale probe selection and chip design of tiling arrays do not depend genome annotation of the sample of DNA under consideration. However, a traditional microarray requires a known annotation, because of the chip characteristics and the experimental objectives that may be possible to obtain using such technology. Tiling arrays get their name from the placing of probes on chip such that they appear to tile the surface. These probes may be arranged on a surface in at least two ways. On approach is place them at random on a surface, and the other is to place them in some predesigned pattern. Among the latter are probes that either overlap, are end to end or are spaced at some predetermined distances to enhance and clarify experimental results. An illustrative diagram of these designs for tiling arrays may be found in the review paper Lang *et al.* (2008) as well as discussions of tiling arrays on many sites on the world wide web, which are too numerous to cite.

To get some idea as to the physical size of microchips, there are now high density chips used in tiling array experiments that hold 6.6 million sites in an area less than 2 cm^2. Therefore, because of the small size of these chips, the production of probes of DNA that are to be placed on these chips, which are essential steps in working with this technology, must be automated, using analytical techniques from chemistry, physics, engineering and computer science along with the writing of software to run the computers. The size of these probes are often expressed on the level of nano-scales. One of the key words that are used in describing structures used in tiling array technology is oligonucleotides, which are short nucleic acid polymers that typically contain fewer than 20 bases. As these small polymers of nucleotides bind to complementary nucleotides, they are often used as probes to detect stretches of DNA or RNA of interest. A practical illustration of the use of this technology is the availability of tiling chips with lengths of 25 to 1000 nucleotides (nt) probes that are usually synthesized, using photolithography or ink-jet technologies as well as other technologies. As these technologies often have capabilities for high throughput and simplicity in the preparation of probes, they are widely used in tiling array experiments designed to elucidate the functions of genes and the location of their regulatory sites on significant stretches of a genome. The paper by Lang *et al.* (2008) may be consulted for a more detailed description of a work flow in the preparation probes for use in tiling array experiments. It should be noted that in preparing probes for tiling array experiments, DNA may be sampled form cells from various tissues. In passing, it is of interest to note that oligonucleotides are also used in a technology called fluorescent in situ hybridization (FISH), which is often used to compare sequences of DNA from different species. This technology was, in fact, used to discover that human and chimpanzee DNA may be matched base by base for about 98 to 99% of their bases.

Although tiling array technology consists of a set of powerful tools for exploring the nature of information encoded in DNA of a genome with relatively fast input and rapid turn around time, the simulated signals generated by tiling arrays are not as reliable as digital information expressed in terms strings of bases of DNA. Consequently, the results obtained from tiling array experiments must be tested and validated using other experimental procedures. Unfortunately however, when searching for sites where regulatory complexes such as transcription factors bind to DNA, the number of putative sites suggested by a tiling array experiment may be so large that it no longer possible to test the validity of such sites using existing procedures. In such cases, mathematical and statistical procedures are often utilized in an effort to reduce the number of sites that need to be tested in further detail. Among the classes models used in this connection are those from information theory as well as hidden semi Markov processes, but detailed description of such mathematical structures are beyond the scope of this chapter.

Readers, who are interested in pursuing statistical methods of analyzing microarray gene expression data further, are advised to consult the book McLachlan *et al.* (2004). Although the aims of the statistical methods described in this book were to applications in the analysis of gene expression data, many of the methods are also applicable to the preparation of probes and the analysis transcription data generated from tiling array experiments.

13.7 A Dispersed View of Genome Activity as Revealed by the ENCODE Project

In this section some of the results obtained by the ENCODE project, using tiling array technology, will be briefly outlined and discussed. More detailed information as well as references may be found in the review paper

Gerstein *et al.* (2007). A first finding of the ENCODE project was that a vast amount of *DNA* not annotated as genes was transcribed into *RNA*, a finding that confirmed earlier exploratory work by other investigators. These novel transcribed regions are called *TARs*, transcription active regions, and transfrags. Although a majority of the human genome appears to be transcribed at the level of primary transcripts, only about half of the spliced transcriptions detected across all cell lines and conditions mapped were currently annotated genes.

A second observation of the ENCODE project was the presence of unannotated and alternative transcription start sites, (TSSs). These sites were identified by either the sequencing of $5^{'}$ end of transcribed *mRNA* or the mapping of promoter associated transcription factors, using *ChIP-chip* or *ChIP-PET* technology. Moreover, the consortium found that many known protein coding genes have alternative *TSSs* that are sometimes $>$ 100 kb upstream from the annotated transcription start site. Experiments were also performed with $5'$ rapid amplification *cDNA* ends (*RACE*) on all 399 well-characterized protein coding loci in the ENCODE regions. The *RACE* primer was selected from a $5^{'}$exon that was shared among most of the annotated transcripts from each locus, and the *RACE* products were hybridized to arrays and mapped. It was found that more than half of the loci had alternative transcription start sites upstream from the known site in at least one of the 12 tissues tested. It was also found that some of the distal *TSSs* used the promoter of an entirely different gene locus. These results were significant, because it was found that some of the *TSS* were two or three gene loci upstream from the locus at which the *RACE* primer was selected. An isoform usually refers to a protein that has some known function as another protein and may also differ in its sequence of amino acids. It was also found by the *ENCODE* project that some alternative isoforms code for the same protein, which differs only in their $5^{'}$ *UTRs.*

This latter finding motivated a research team at the Sanger Institute in the *UK* to produce a detailed annotation called *GENCODE.* In this connection, this team found that the number of protein-coding loci had not increased, but, on the other hand, the number of annotated isoforms had increased. The *ENCODE* project has also provided evidence for dispersed regulatory regions throughout the genome that appear not to occur in a random manner. Some regions of the genome were rich in regulatory sites; whereas in some regions these sites were sparse. Another finding of the *ENCODE* project was that the terms, genic and intergenic, become problematic. For according to the traditional view, genes are unitary regions of a

DNA sequence separated form each other, but a finding of the *ENCODE* project was that if one attempts to define genes on the basis of shared transcripts, then many annotated distinct loci coalesce into larger genomic regions. Furthermore, there is much more activity between annotated genes in the intergenic space than was previously thought. This activity is characterized by non-coding proteins genes that transcribe *ncRNA* and pseudogenes in the intergenic space. It also appears that some transcribed elements are under evolutionary constraint, *i.e.*, they are "held" there by evolutionary forces such a natural selection. Interestingly, some *ncRNA* genes and pseudogenes are located within introns of protein-coding genes. Pseudogenes may bear close resemblance to known genes at other loci, but have been rendered non-functional by additions or deletions in structure that prevent normal transcription and translation. When attempting to define a gene, the complications just outlined should not be overlooked.

As mentioned in a foregoing section, there are *ncRNA* that have been transcribed and it has been recognized that the roles *ncRNA* genes play are quite diverse. Included in these roles are gene regulation, *miRNAs*, *RNA* processing, *snoRNA*, and protein synthesis, *tRNA* and *rRNA*. As these genes lack codons and open reading frames, they are difficult to identify, and, thus it is probable that only a fraction of such genes in the human genome have been identified as of 2009. There are exceptions, however, for the case of some *ncRNAs*, which can be identified computationally through *RNA* folding and coevolution analysis. Examples of such *RNAs* include *miRNAs* that display characteristic hairpin-shaped precursor structures and *ncRNAs* in ribonucleoprotein complexes that in combination with peptides form specific secondary structures. An example of a *ncRNA* gene is the *XIST* gene, which acts functionally to inactivate one of a pair of *X* chromosomes during the early stages of development of mammalian females. This inactivation of one *X* chromosome provides dosage equivalence for males and females, when males have only one *X* chromosome. There is evidence that a transcript of this gene is spliced but apparently does not code for a protein. If the phrase "*XIST* gene" is entered into a search engine for the world wide web, much more information on this gene may be obtained. The number of bases is this gene has been estimated to be 17-*kb*, which shows that functional *ncRNAs* can expand significantly beyond the size of a constrained computationally identifiable region of *DNA*.

Another group of mysterious genomic structures that are often found in introns and intergenic space are pseudogenes, which were defined above. Apparently, some of these genes "move" among the states of being active or non-active and can thus influence the structure and function of a human genome. It has been reported that number of pseudogenes is about equal to protein coding genes, and, therefore, they have a confounding effect on gene annotation. It has also been reported recently, that as many as 20% of these genes are transcriptionally active, which suggests that caution should be exercised when using expression as evidence for locating genes. If a reader is interested in finding more information on these types of genes, the phrase "pseudogenes" should be entered into a search engine for the world wide web. Indeed, there are web sites that are devoted to pseudogenes.

The term, constrained elements, refers to sections of a genome that apparently have been held there by evolutionary forces. It has been found by that by comparing evolutionary changes and similarities in genomes of multiple species and within human populations that non-coding regions of genomes contain a large fraction of functional elements. Workers in the *ENCODE* project found that only about 40% of the evolutionary constrained bases lie within protein coding exons and their associated untranslated regions. The resolution of constrained bases among species and individuals is high, for in some cases these regions contain as few as 8 bases, and the median number of bases in such regulatory sites has been reported to be 19. Such observations suggest that protein coding loci may be viewed as clusters of evolutionary constrained *DNA* sequences dispersed within a sea on unconstrained sequences. It has also been found that approximately 20% on the evolutionary constrained elements are located in protein coding regions as unannotated non-coding regions, which suggest that non-coding regions may be as functionally important as coding regions. Just as with other observations mentioned in this section, these findings should be taken into account when attempting to define a gene in terms of segments of *DNA*.

As illustrated above, much biological complexity was revealed by the ENCODE project. Contained in the remainder of this section are some symbolic hypothetical examples of the variety of *RNA* transcripts found by the ENCODE project. The following table is a typical representation of a genomic region containing four genes, which are similar to regions revealed by the ENCODE project.

$5'$	Gene 1	·	TAR 1	·	Gene 2	·	TAR 2	Gene 4	·	$3'$
$3'$	·	·	·	·	·	·	Gene 3	·	TAR 3	$5'$

In this table the symbol Gene denotes a set of contiguous bases which contain protein coding regions called exons and non-coding regions called introns. The symbol $\cdot$ denotes a sequence of bases among genes and transcription active regions, TAR. As in previous sections, the upper strand is the sense strand and the lower is the antisense strand. A symbolic representation of Gene 1 has the form

$$\text{Gene 1} = \begin{array}{|c|c|c|c|c|c|c|c|c|}\hline \alpha & \cdot\cdot & \beta & \cdot\cdot & \gamma & \cdot\cdot & \delta & \cdot\cdot & \epsilon \\ \hline \end{array},$$

where the lower case Greek letters stand for annotated exons containing various numbers of bases and the symbol $\cdot\cdot$ stands for annotated exons with varying numbers of bases. In what follows the symbol $\cdot\cdot$ stands for annotated introns and the lower case Greek letters represent annotated exons. Given this system of notation, in the genomic region under consideration, Gene 2 has the symbolic form

$$\text{Gene 2} = \begin{array}{|c|c|c|c|c|c|c|}\hline \varepsilon & \cdot\cdot & \zeta & \cdot\cdot & \eta & \cdot\cdot & \theta \\ \hline \end{array},$$

Gene 3 has the symbolic form

$$\text{Gene 3} = \begin{array}{|c|c|c|c|c|c|c|}\hline \vartheta & \cdot\cdot & \iota & \cdot\cdot & \kappa & \cdot\cdot & \lambda \\ \hline \end{array}$$

and Gene 4 has the symbolic form

$$\text{Gene 4} = \begin{array}{|c|c|c|c|c|c|c|c|c|}\hline \mu & \cdot\cdot & \text{TAR 4} & \cdot\cdot & \nu & \cdot\cdot & \xi & \cdot\cdot & \pi \\ \hline \end{array}$$

The ENCODE project also revealed a variety of RNA transcripts starting the $3^{'}$ end of the antisense strand of the DNA molecule symbolized above. Among the transcripts of the antisense strand was a strand of the form

$$\begin{array}{||c||c||c||c||c||c||c||c||c||c||c||}\hline\hline 5^{'} & \alpha & \cdot\cdot & \beta & \cdot\cdot & \gamma & \cdot\cdot & \delta & \cdot\cdot & \epsilon & 3^{'} \\ \hline\hline \end{array},$$

which is Gene 1 including the introns. Just as in previous sections, the double bar reminds the reader that the table represent a molecule of RNA rather then DNA. Another RNA transcript had the form

$$\begin{array}{||c||c||c||c||c||c||c||c||c||c||c||c||c||c||}\hline\hline 5^{'} & \alpha & \cdot\cdot & \text{TAR 1} & \cdot\cdot & \cdot\cdot & \varepsilon & \cdot\cdot & \zeta & \cdot\cdot & \eta & \cdot\cdot & \theta & 3^{'} \\ \hline\hline \end{array},$$

which included the exon α from Gene 1, TAR 1 and all of Gene 2. Also among the transcripts was a RNA strand of the form

$$\begin{array}{||c||c||c||c||c||c||c||c||c||c||c||c||c||}\hline\hline 5^{'} & \alpha & \cdot\cdot & \text{TAR 2} & \cdot\cdot & \mu & \cdot\cdot & \text{TAR 4} & \cdot\cdot & \nu & \cdot\cdot & \pi & 3^{'} \\ \hline\hline \end{array},$$

Observe that the exons in this transcript are composed of one exon α from Gene 1 and the exons μ, ν, π and TAR 4 from Gene 4. Also contained in this transcript is the region TAR 2 between genes 2 and 4 on the sense strand of the DNA molecule.

There were also transcripts of the sense strand of the DNA molecule, starting from the $3^{'}$ end, which led to, among other things, transcripts of Gene 3. Sometimes, transcripts of the sense strand are also called reverse transcripts. Among the reversed transcripts was one of the form

$$\begin{array}{||c||c||c||c||c||c||c||c||c||c||c||}\hline\hline 5^{'} & \vartheta & \cdot\cdot & \iota & \cdot\cdot & \kappa & \cdot\cdot & \lambda & \cdot\cdot & \text{TAR 5} & 3^{'} \\ \hline\hline \end{array},$$

which contains all the Gene 3 plus the region TAR 5.

In this hypothetical example, regulatory sequences of Gene 1 occurred close to Gene 1 and also in Genes 2 and 4 as symbolized in the DNA strand below

$$\begin{array}{|c|c|c|c|c|c|c|c|c|c|c|}\hline 5^{'} & \star & \text{Gene 1} & \star & \text{TAR 1} & \star & \text{Gene 2} & \star & \text{TAR 2} & \text{Gene 4} & \star & 3^{'} \\ \hline \end{array},$$

where the symbol $\star$ stands for a regulatory sequence for Gene 1. In terms of base pairs, the distances among these regulatory sequences can be quite large. In this example, it was also supposed that Gene 2 contained a regulatory sequence $\star$ for Gene 1 as denoted below

$$\text{Gene 2} = \begin{array}{|c|c|c|c|c|c|c|c|}\hline \star & \varepsilon & \cdot\cdot & \zeta & \cdot\cdot & \eta & \cdot\cdot & \theta \\ \hline \end{array}\ .$$

It was also assumed that Gene 4 contained three regulatory sequences $\star$ for Gene 1 as symbolized below

$$\text{Gene 4} \ = \begin{array}{|c|c|c|c|c|c|c|c|c|c|c|c|}\hline \star & \mu & \cdot\cdot & \text{TAR 4} & \star & \nu & \cdot\cdot & \xi & \cdot\cdot & \pi & \star \\ \hline \end{array}\ .$$

Among the transcripts of Gene 1 from this type of configuration of DNA were those of the form

$$\begin{array}{||c||c||c||c||c||c||c||c||c||c||c||c||c||c||}\hline\hline 5^{'} & \cdot\cdot & \cdot\cdot & \cdot\cdot & \text{TAR 2} & \cdot\cdot & \mu & \cdot\cdot & \text{TAR 4} & \cdot\cdot & \nu & \cdot\cdot & \pi & 3^{'} \\ \hline\hline \end{array}.$$

It is of interest to note that this transcript contains no exons form Gene 1, the exon ξ is missing from Gene 4, and that none of the regulatory sequences, symbolized by $\star$, were contained in this transcript. Another transcript of Gene 1 had the form

$$\begin{array}{||c||c||c||c||c||c||c||c||c||c||c||c||c||c||}\hline\hline 5^{'} & \text{TAR2} & \cdot\cdot & \mu & \cdot\cdot & \text{TAR 4} & \cdot\cdot & \nu & \cdot\cdot & \xi & \cdot\cdot & \pi & 3^{'} \\ \hline\hline \end{array}.$$

Observe that this transcript had all Gene 4 but no regulatory sequences $\star$. In this example, some regulatory sequences for Gene 1 are actually closer to other genes and would be misidentified if they were chosen on the basis proximity.

The illustrative model of genes and transcripts displayed above typify the transcripts observed by the workers of the $ENCODE$ project in that they span multiple loci with some using distal $5^{'}$ transcription start sites. Many novel isoforms were also identified among the transcripts of the genomic regions of the form displayed above.

13.8 Statistical Models as a Necessary Basis for Interpreting Data From Tiling Array Experiments

Recently, a large amount of DNA transcription data has been generated using high density tiling array microchips, see Gerstein *et al.* (2007) for references. An advantage of using such technology is that it provides a means for constructing probes to detect DNA transcription in a detailed and unbiased way without preconceptions as to where to look for such activity. The interpretation of output signals from tiling array experiments needs however, to be carefully analyzed to result in a description of a reliable set of regions of DNA that undergo the process of transcription. However, the amount of detected transcription depends significantly on thresholds used in the designation transcribed regions and also on the segmentation algorithms used in the segmentation of DNA into transcribing and non-transcribing regions. An interesting paper on the use of Bayesian methods in DNA sequence segmentation is that of Boys and Henderson (2004). Moreover, because the experiments to detect transcription are carried out on many different tissues and cell lines, the direct comparison among experiments is not an easy task and is sometimes further complicated by small overlaps of transcription regions from separate experiments. Such complications may be due, in part, to the intrinsic background variability among the biological materials used in the experiments.

Transcription mapping refers to a process of attempting to find those regions of DNA or RNA that undergo a transcription process, resulting in the production of transcription products such as the various types of RNA discussed in previous sections. In most transcription mapping experiments, the actual transcription map is not known, which gives rise to problems of dealing with uncertainty in identifying transcription regions. One approach to assessing the errors that may arise in the identification of such regions is to randomize the original data and then apply scoring, segmentation and other schemes to detect transcription regions in this seemingly meaningless data. Such procedures may yield estimates of the probabilities of false positives in "null' models constructed by a randomization procedures, using, among other things, computational algorithms similar to those discussed in section 5.7. By generating a sample of such experiments, one could estimate the distribution of classification errors in the null case. It is not clear however, as to what sense such procedures are optimal when other ways of assessing uncertainty in the identification of transcribing regions are taken into account. Some factors that may contribute to difficulties in using such procedures include GC content of a region of DNA or RNA, actual or expected lengths of transcription regions and dinucleotide composition. The expected outcome of such experiments also depend on the tissue or cell line being used, developmental stage of materials and a host of other factors. As a means of working toward greater confidence in predicted regions of transcription, comparison of verified transcription maps from other experiments can also be helpful.

For those readers who are interested in learning about methods of scoring used in DNA sequence and genome analysis, the book, Mount (2002), which was written for biologists, may be consulted. Included in this book are, for example, chapters on quantitative and graphical methods used in predicting RNA secondary structures and the scoring techniques used in predicting those regions of DNA corresponding to genes. It would also be of interest to consult the book on statistics by McLachlan *et al.* (2004) for methods of analyzing gene expression data generated by using gene micro-chip technology.

Various methods based on fields in mathematics, probability and statistics have been proposed and used in analyzing data on DNA transcription. Among these methods is that of splicing graphs introduced by Heber *et al.* (2002). Traditional approaches used in the annotation of alternative splicing is to investigate every splicing variant of a region of DNA containing introns and exons (gene) on a case-by-case basis. Recent studies have shown that alternative splicing of a gene occurs more frequently than previously thought, which gives rise to a need for procedures that can systematically describe these events in a coherent way. These authors introduced the notion of the splicing graph, which is a natural and useful way to represent all splicing variants of a gene. To accomplish this goal they replaced the linear sequence representation of each transcript with a graphical representation such that each transcript corresponds to a path in the graph. The authors also designed an algorithm to assemble EST, expressed sequence tab, reads into the splicing graph rather than describing each variant on a case-by-case basis. A web site referenced in the paper may be consulted for the software implementing this algorithm.

Also among the collection of statistical models used in analyzing data generated by tiling array experiments are those referred to as hidden Markov models. A technical description of this class of models will not be given here but the interested reader may consult the recent book Barbu and Limnois (2008) for detailed discussions. The $p53$ gene is a tumor suppressor gene, which acts to regulate the cell cycle and has been mapped to chromosome 17 of the human genome. The expressed form of this gene is a protein labeled $p53$. As mentioned previously Transcription factors, TFs, regulate gene expression by binding to specific regulatory regions in a genome, which in eukaryotes may occur long distances away from the regulated gene. Recent advances in technology have led to the development of high-density oligonucleotide arrays that tile all non-repetitive sequences of the human genome up to 35 *bp* resolution. Array platforms of this type allow for the unbiased mapping of in vivo TF binding sequences, $TFBS$, using techniques described briefly in Li *et al.* (2005), who used a hidden Markov model (HMM), among other things, to locate $TFBSs$. Also used in their analysis was a Motif Discovery scan algorithm, $MDscan$, to recover the $p53$ motif from $p53$ target regions identified by using the HMM. Software implementing this algorithm is available on the web. Furthermore, it was found that several of the newly identified $p53$ $TFBSs$ are in the promoter regions of known genes or associated with previously characterized $p53$-responsive genes. In passing, it is of interest to note that the $p53$ gene codes for a protein that is also a transcription factor involved in the regulation of other genes implicated in regulating the cell cycle.

Before proceeding to outline the contents of recent papers on applications of hidden Markov in segmenting DNA, it will be helpful to cite a few related papers form the statistical literature. Even though the interesting paper by Jensen *et al.* (2004) on the computational discovery of gene regulatory binding motifs is written form a Bayesian perspective, it also contains non-Bayesian material on the use of the concept of entropy and related concepts in exploring samples of DNA for a sets of sites where gene regulatory elements bind. Other papers from the statistical literature on applying hidden Markov models in the statistical analysis of DNA samples are those of Boys *et al.* (2000), (2002) and (2004), which deal, respectively, with detecting homogeneous segments in DNA sequences, and determining the order of Markov dependence in an observed process and a Bayesian approach to DNA segmentation. Of particular interest was the paper Boys *et al.* (2002) which deals with the problem of determining the order of Markov dependence on an observed process of a sequence of bases, which are assumed to have evolved as a hidden Markov model. In some papers, for example, Jensen *et al.* (2004) it is assumed that bases making a sequence of DNA and statistically independent.

Du *et al.* (2006) have used a set of methods for efficiently segmenting tiling array data in transcriptional and *chIP-chip* experiments described in a previous section. These authors used what they termed a supervised hidden Markov model framework that systematically incorporates validated biological knowledge. To illustrate some on the ideas described by these authors, let D denote a region of DNA under investigation and let $(U_1, U_2, \ldots, U_m)$ denote a collection of m disjoint subregions of D such that each subregion is of $k \geq 1$ base pairs in length. Some authors have also considered sampling schemes in the disjoint regions of DNA do not have the same number of base pairs. Briefly, these authors use a statistical measure based on concept from entropy and information theory to explore these disjoint regions for preliminary evidence of the process of transcription, and this information is incorporated into hidden Markov models for a more refined exploration of regions of DNA that undergo the process of transcription. The idea of learning processes were also used by these authors in the sense that if m is an even number, then the information gleaned from exploring a sample of $m/2$ disjoint regions is used in an attempted confirmatory way by checking whether the predictions resulting from the exploratory sample are confirmed in the remaining sample of $m/2$ disjoint regions of DNA.

As an another example of using a heterogeneous hidden Markov model for segmenting Comparative Genomic Hybridization (CGH) data the paper by Marioni *et al.* (2006) may be consulted. Presented in this paper is a method for segmenting such hybridization data into states with the same copy number and the process is heterogeneous in the sense that the transition matrix-valued function depends on the median of the copy number of the preceding state. The few papers cited in this section are merely a sample from a variety of papers published in various mathematical and statistical journals including Bioinformatics, Statistical Science, The Annals of Applied Statistics, Biometrics and others.

13.9 A Tentative Updated Definition of a Gene

Throughout the preceding 12 chapters of this book the idea of a gene has been treated at the Mendelian level as an abstract object with alternative forms, alleles, upon which the components of natural selection may act. Furthermore, it was assumed that mutations could occur such that one allele could mutate to another. In the foregoing sections of this chapter, evidence has been cited that calls the central dogma of biology into question, because there are transcription products, $RNAs$, that do not code for proteins but may have other functions such as acting as catalysts for chemical reactions which are necessary for the proper functioning of a cell. Such observations compel researchers to consider a more general definition of a gene that provides a framework for a broader understanding at to what constitutes a gene at the molecular level. Such an updated definition should also be capable of aiding in interpretation of experimental results obtained by using recently developed technology such as tiling arrays.

To start the discussion a gene will be defined as a genomic sequence consisting of either DNA or RNA that directly codes for functional product molecules consisting of either protein or RNA. In this definition, RNA was included in the genomic sequence, because it was desired that the concept of a gene should also include such entities as some viruses whose genomes consist only of RNA. For the moment, the term, functional product molecules, will not be defined but will be discussed subsequently. It should also be mentioned that, like some of the foregoing sections of this chapter, the thoughts recorded in this section were inspired by the paper Gerstein *et al.* (2007).

By way of further clarification of the definition, different functional products, protein or RNA, that were encoded by the same genomic region, consisting of disjoint exons, for example, but may have been derived from different subsets of exons or subsets of the region, are included in the same gene. To experimentally determine these subsets of a region of DNA or RNA which code for a set of functional products, either amino acid or RNA sequences, one would project downward onto the original genomic sequence from which it was derived or coded. Given a genomic sequence, one could also, in principle, project upward to a set of functional products that were coded for by subsets of the genomic region, but, because it is a common practice to annotate genomic regions, it seems preferable to project downward from a set of functional products in attempts to determine genomic regions responsible for the coding of these products. It should also be stated that when looking for genomic products with common sequence segments, sequence identity is not enough, *i.e.*, they must be coded by the subsets of the genomic region under consideration. In this connection, it is of interest to consider that case of paralogous proteins which are coded for by duplicate genes, which are thought to arise from unequal crossing over during meiosis. Such proteins can have sequences in common but are coded for by different DNA sequences in different of parts of the genome. Hence, by definition, these separate sequences would not constitute one gene. There is an extensive literature on the evolution of paralogous proteins and for further information an interested reader may wish to consult sites on the world wide web by entering this term into a search engine.

In those cases in which there is a set of functional products encoded by the genomic region under consideration, a gene will be defined as the union of all those sets of bases such that the functional products can be projected downward onto them. It will be instructive to sate this part of the definition of a gene in more formal terms. Let GR denote the genomic region under consideration and let FP denote the set of functional products encoded by sets of bases in GR. For every $f \in FP$, let $A_f \subset GR$ denote the set of bases that code for the product f. Then, the gene, denoted by *Gene*, in the region GR will be defined as the union

$$Gene = \bigcup_{f \in FP} A_f. \tag{13.9.1}$$

This definition is coherent in the sense that $A_f \subset GR$ for every $f \in FP$. It also follows from this definition that $Gene \subset GR$ and in some cases it may happen that $Gene = GR$. The sets of bases in the union (13.9.1) may or may not be disjoint, but in any case this union has the property that if the set FP contained only one product f that mapped onto the set A_f of bases, then the union would reduce to the set A_f. Observe that this definition specializes to earlier definitions of a gene in which, for example, a protein may be encoded for by a continuous set of bases.

The are cases in which a genomic region of DNA contains a set of exons such that *pre-mRNA* can be alternatively spliced to generate a $mRNA$ with a frameshift that encodes for a different protein. An interested reader may consult Gerstein *et al.* (2007) for a reference on this phenomenon. In these cases two *mRNAs* may have sequences in common but code for proteins with entirely different properties. This unusual case raises the question as to how exactly sequence identity is to be handled when taking the union of sequence segments shared among protein products. If one considers the set of protein products in such situations, there must be two genes with overlapping or non-disjoint set of bases that code for the proteins. If one projects these sequences onto the sets of bases that encode them in some genomic region GR however, then there are two sequences with non-disjoint set of bases from which the were encoded. Thus, according to the definition in (13.9.1), there is one gene with sets A_f and A_{f^*} in GR with common elements such that $A_f \cap A_{f^*} \neq \varphi$. If there is evidence that these two sequences have been constrained by evolution such that a mutation in either A_f or A_{f^*} affects the two proteins simultaneously, then this would suggest that this situation is entirely different from cases in which there are two unrelated protein coding genes under consideration. Generalizing from this special case seems to support the idea of taking a union of coding sets as in (13.9.1) is a useful way to define a gene. It is of interest to note that those cases in which a genomic region contains many exons which give rise to the phenomenon of alternate splicing are also included in the definition of a gene exemplified in the union (13.9.1).

Gerstein *et al.* (2007) suggest that regulatory regions should not be considered when deciding whether multiple functional products are encoded by the same gene, even though these regions play an important role in gene expression. This facet of the definition of the gene under consideration stems from the authors's concept of a bacterial operon. Traditionally, the fact that gene in the operon share an operator and promoter region has been interpreted that their protein products are alternative products of the same gene. For the case of higher eukaryotes, however, the situation seems to be more complicated. This is because in this class of organisms, it has been observed that there may be two transcripts that originate from the same transcription start site and share the same promoter and regulatory regions but their functional products do not have common elements, which may be due to alternative splicing. For the case of proteins, these products would have different sequences of amino acids. In such cases, the authors would not consider that products of this type as being coded for by the same gene, but were coded by two different genes. The authors concluded that regulation is simply too complex to be included in the concept of a gene, when attention is being focused on mapping products of transcription onto some genomic region.

Nevertheless, there is evidence that mutations in regulatory regions have played significant roles in evolution. For an account of this evidence for the general reader, the interesting paper Carroll *et al.* (2008) may be consulted. Contained in this paper are accounts as to how changes in regulatory regions of the DNA account for morphological changes among species which share many genes but differ in terms of regulatory switches that turn these genes off and on. Of interest on how a mutation in regulatory regions affect human biology is that of the Duffy protein. This protein usually appears on the surface of human red blood cells but also functions in the brain, spleen and kidneys with separate enhancer sequences of DNA for each of these functions. On red blood cells, this protein forms part of a receptor that the malarial parasite, *Plasmodium vivax*, uses to enter the cell. Nearly 100 percent of West African populations lack Duffy proteins on their red blood cells, which renders them resistant to malarial infection. Interestingly, the

Duffy gene's red cell enhancer in these individuals has been disabled by a mutation in a single letter of the DNA sequence. In particular, this mutation has the form $T \to C$, which is a concrete example of a single nucleotide substitution turning off an enhancer, but, interestingly, the other Duffy enhancers are not affected. A developing field of evolutionary study in which attention is focused on regulatory regions of DNA implicated in turning genes off and on is referred to as evolution and development and is denoted by acronym *evo devo.* Because of known cases similar to that just described, Gerstein *et al.* (2007) suggest that genomic regions that play important roles in gene expression be called gene-associated regions, which may play very significant roles in evolution.

A basic aspect of the proposed definition of a gene is the requirement that the protein or RNA must be functional when they are assigned to a gene. This aspect of the definition preserves a basic principle of genetics that genotype determines phenotype in connection with interactions with the environment. At the molecular level phenotypes may be defined in terms of biochemical function, which seems to be compatible with earlier concepts of a gene. Given these comments, the next step is to move to the question "what is function?". In finding answers to this question, it seems highly likely that high-through put biochemical and mutational assays will be needed to define function on a large scale, and, it is hoped that it is only a matter of time until experimental evidence is obtained that will establish what most proteins and $RNAs$ do.

On the other hand, it may be the case that science will not be able to ever know the function of all molecules in a genome, and, moreover, it is conceivable that some genomic products are just noise and are the results of evolutionary neutral events which are tolerated by organisms. In this connection, it is very interesting to note that there is a great deal research activity directed towards understanding the roles $RNAs$ and proteins play in attempts to gain an understanding of the function of these products. If, for example, one enters the phrase "regulatory RNA" into a search engine on the internet one is immediately led to hundreds sites that either report the results of recent research or are thoughtful discussions as to the roles these products play human biology and that of other organisms. It is also interesting to note that there is a substantial amount of research activity directed to finding gene-associated regions that play basic roles in gene expression. Such regions are also known as the epigenome and are of basic importance in the emerging field of epigentics. Recently, it has been suggested that comparative genomics may in future play a significant role in finding gene-associated regions, see Carroll *et al.* (2008) for details. An interesting example of this situation is the case that, although it has been estimated that sharks and man shared a common ancestor about 500 million years ago, it has been observed that sharks and man share nearly 5,000 non-coding genomic regions that appear to be enhancers, which are near genes involved in body building. Evidently, such proximity, which has been preserved by natural selection, is a reflection of the overall body architecture shared by all vertebrates. Examples of this kind may one day may play a significant role in finding conserved genomic regulatory regions that make up the epigenome humans and other species.

13.10 Genetics of ABO Blood Group in Humans

In biology one is compelled to give specific examples of phenomena rather than discussing them in general terms. In this and subsequent sections of this chapter, specific examples of the genes and their regulatory systems will be given. It has often been stated that the ABO blood system in humans is the most important with respect to the safe administration of blood transfusions. Broadly speaking, there are four phenotypic blood groups, which are classified as A, B, AB and O. At the molecular level, these blood groups are distinguished by the antigens found in an individual's blood. It is of interest to note that these blood groups are also found in species who are thought to share common ancestors with man such as chimpanzees, bonobos and gorillas. From the point of view of Mendelian genetics, the four blood groups are the result of the actions of three alleles at a locus located on human chromosome 9. In what follows, an overview of the alleles at the molecular level will be given.

A chemical structure called the H antigen is an essential precursor to the ABO blood group antigens. It has been found that the H locus, genomic region, is located on chromosome 19 and contains 3 exons that span more than 5 kilo bases, kb, of genomic DNA. It is thought that these exons, when linked together, encode a chemical called fucosyltransferase that produces the H antigen on the red blood cells. Briefly, the H antigen is a carbohydrate sequence linked mainly to a protein structure. From the point of view of classical genetics, we shall see below that the expression of the A, B, O phenotypes involve the interaction of two loci located of different chromosomes.

The ABO genomic region is on chromosome 9 and contains 7 exons that span more than 18 kb of genomic DNA. The exons range in size from 28 to 688 base pairs, bp, and, the two largest exons, 6 and 7, encode most of the coding sequence, which is 1062 bp in length. The three alleles, A, B and O, can be characterized to some extent at the genomic and molecular levels. From the point of view of the general reader, the A allele encodes a glycosyltransferase, an enzyme, that bonds a chemical to the D-galactose end of the H antigen, producing the A antigen. The B allele encodes a glycosyltransferase that joins to still another chemical structure which bonds to D-galactose end of the H antigen, thus creating the B antigen. For the case of the O allele, exon 6 contains a deletion that results in a loss of enzymatic activity. The O allele differs from the A allele by the deletion of a single nucleotide, Guanine at position 261, which leads to a reading frameshift that give rise to an almost entirely different protein that lacks enzymatic activity. As a result of this shift, the H antigen on the red blood cells remains unchanged for the case of the O allele.

From an evolutionary perspective, mutations in human genomic DNA involved in the regulation of genes can have significant impacts on the survivability of infants and adults. As one searches the internet for information on the regulation of genes or reads papers on of genes involved in development, it becomes apparent that the terms promoters and enhancers are frequently used. Promoters are chemical structures that bind to regions of DNA near a transcription start site and have been implicated in the initiation of transcription. Enhancers, on the other hand, are chemical structures that bind to regions of genomic DNA that are often at considerable distances from a gene, as expressed in terms of the number of bases pairs, and have been implicated in qualitative and quantitative aspects of gene expression. Readers wishing to get an overview of research activity on promoters involved in the human H locus on chromosome 19, for example, one can type the phrase "human H locus promoters chromosome 19" into a search engine, which often results in immediate access to scholarly articles on a subject. Another example of a related phrase is "H gene promoter region human chromosome 19", which leads to scholarly articles on the location promoter regions for the H gene. It is beyond the scope of this section to discuss the large array of articles related to these subjects, but suffice it to say the H locus seems to code for chemical structures that have been implicated in the search for genomic regions of DNA affecting various autoimmune human diseases. In the language of classical genetics, the human H locus on chromosome 19 has many pleiotropic effects. Essentially the same comments apply when the phrase "human *H* locus enhancers chromosome 19" and related phrases are entered into a search engine, but in these cases, the enhancer regions are often distant from the transcription start site of a gene.

With regard to the *ABO* locus on human chromosome 9, access to scholarly papers on a promoter region may be obtained by entering a phrase of the form "*HBO* gene promoter region human chromosome 9" into a search engine. An example of the papers on this subject is that of Kominato *et al.* (2002), see nlm.nih.gov/pubmed/11856466. In this paper, among other things, the authors provide information on the location of the promoter region for this gene and also on how a mutation in this region abrogates the binding of a transcription factor with the result that *ABO* promoter does not function properly in erytholeukaemina and gastric cancer cells.

Access to scholarly papers on enhancers may be obtained by entering the phrase "*ABO* Enhancers" into a search engine. In this connection, the interesting paper Yu *et al.* (2000) may be consulted, see nlm.nih.gov/pubmed/10873628. These authors demonstrated that the enhancer for the *ABO* glycosyltransferase gene was about 3.7 kb upstream form the transcription start site and that is was composed of four tandem repeats of a 43 bp (base pairs) unit. Moreover, these authors demonstrated that the enhancer structures differ among the alleles A, B and O of the glycosyltransferase genes. The enhancer with four 43 bp units is present in both the B and O genes. However, the corresponding enhancer for the A gene contains only one 43 bp unit, and within this unit, a nucleotide substitution was shown to exists by comparing it with the consensus sequence of the 43 bp unit. The differences in the repeat numbers of the 43 bp unit two allelic genes was shown to be the principal reason for the vast differences in the transcriptional activities between the A-gene and B-gene enhancers.

As indicated the paper just discussed was published in 2000, but in more recent years, it has been found that copy number of some units of *DNA* varies among genomes of individuals and some have been implicated in human diseases. For more information on this developing field of research an interested reader may enter the acronym *CNV* (Copy Number Variation) into a search engine on the internet. One such search performed by the authors, for example, yielded about 2,810,000 hits, which is emblematic of the amount of research activity in this field.

13.11 Duffy Blood Group System in Man

The gene locus for the Duffy blood group system is referred in the literature by the acronym $DARC$ (Duffy antigen receptor for chemokines) and it located on chromosome 1, see Pulst (1999) for more details. Interestingly, the location of this locus was determined by using classical linkage association techniques. For this system, the antigenic determinants reside in an acidic glycoprotein symbolized by (gp-$DARC$ or gp-FY). This structure spans a cell membrane seven times and has an exocellular N-Therminal domain and an endocellular C-Terminal domain. The transcription unit of the Duffy locus includes 1572 nucleotides, which include exon 1 with 55 nucleotides, a single intron of 479 nucleotides and exon 2 consisting of 1038 nucleotides. When compared to the ABO locus on chromosome 9, the Duffy locus contains relatively few nucleotides. A most notable property of this glycoprotein is that it is a receptor for the human malaria parasite, Plasmodium vivax. There is also an orthologous mouse Duffy gene symbolized by Dfy but no further details will be given here.

With regard to function, it has been observed that the protein gp-FY plays a role in inflammation and in malarial infection, and in addition to being a receptor for the human malarial parasite it is also a receptor for the simian malarial parasite Plasmodium knowlesi. The parasite specific binding site, the binding site for the chemokines and the major antigenic domains are located in overlapping regions of the exocellular N-Terminal terminus. With regard to tissue distribution, gp-FY is expressed in erythroid and non-erythroid cells, which include endothelial cells of capillary and postcapillary venules, the epithelial cells of kidney collection ducts, in lung alveoli and in the Purkinje cells of the cerebellum. At present, there is no apparent disease association in humans but it has been suggested that the lack of $DRAC$ protein on the surfaces of red blood cells is associated with susceptibility to $HIV/AIDS$. More information on blood group systems in man may be obtained by entering the acronym $BGMUT$ into a search engine for the internet. Currently, 40 genes and 1128 alleles have been described and reported on this internet source maintained by the NIH, the US national institutes of health, and a multitude of researchers.

The Duffy blood group system is defined in terms of three alleles, which are common in human populations. Two alleles, symbolized by FYA and FYB, encode for two antithetical antigens denoted by Fy^a and Fy^b. An allele denoted by the symbol FYB^{ES}, where ES denotes erythroid silent, is a major allele in African Americans and also in populations of Blacks in Africa but rarely in other populations. The mechanism underlying this allele is a mutation in the $GATA$ motif in the promoter region which abolishes expression of *gp*-FY on erythroid but not in non-erythroid cells. This mutation is a single nucleotide substitution in the third position of the nucleotide sequence $GATA$ such that the mutation $T \to C$ leads to the sequence $GACA$. The reference Tournamille *et al.* (1995) may be consulted for details. Observe that this mutation was also discussed by Carroll *et al.* 2008. If the phrase "$GATA$ Motif" is entered into a search engine for the internet, then one is also led to the phrase "$GATA$ transcription factor". $GATA$ transcription factors are a family of transcription factors that are characterized by their ability to bind to the nucleotide sequence $GATA$, which occurs in many species whose DNA gene sequences have been investigated.

There are also a number of minor alleles at the Duffy locus on human chromosome 1. If an investigator wishes to search for these minor alleles, one may use the DNA alteration if it is known. On the other hand, one may also search by phenotype or by the designation used by an author. In this connection, the word "alias" is often used. If a reader is interested, the NIH web site for the Duffy gene may be consulted in which a table naming the alias, the nucleotide change, the phenotypes and the amino acid change is displayed. Altogether 11 alleles are listed in this table and it is interesting to observe that some nucleotide substitutions result in a change in the coding for an amino acid in a protein, while others are in the regulatory regions for the Duffy locus. These observations point to the conclusion that if one is interesting in studying the effects of mutations form an evolutionary perspective, attention will need to be focused not only on mutations in those part of genes that code for proteins but also those in regulatory regions that may, among other things, affect the expression of transcription products.

13.12 Regulation of the Shh Locus in Mice by an Enhancer 1 Mb Upstream as a Conceptual Model

When one is concerned with either simulating the evolution of a genomic region through the process of mutation and selection or the statistical analysis of observations on this region, it is prudent to study specific examples of genomic regions that have been elucidated at the level of nucleotides so that it is possible to get some preliminary estimates of the number of base pairs to be considered. An example of such a genomic region is that for the sonic hedgehog gene, *Shh*, which is located on chromosome 5 of the mouse genome. This example is of particular interest because its expression in mouse limb buds of a developing embryo is regulated by a long-range enhancer 1 Mb upstream of the *Shh* promoter, see the paper by Amano *et al.* (2009) for details. With this information in view, an investigator knows that a computer model of a genomic region under consideration must contain at least 1 Mb or 10^6 base pairs. To this number one must also add, at a minimum, an estimate of the number of base pairs making up the region encompassing gene *Shh*. By consulting Figure 1 A of Amano *et al.* (2009), it can be seen that a genomic region denoted by $Lmbr1$, limb region 1, is also on chromosome 5 and contains a regulatory region denoted by $MFCS1$, which stands for mammal-fish conserved-sequence 1. Another aim of this section is to provide a brief review of the technical difficulties encounters when investigators attempt to find DNA on the same or other chromosomes where proteins and other chemical structures bind to effect the expression of a genomic region or gene in question.

Thanks to the large amount of activity on sequencing the genomes of human and other species by various groups of scientists, much information is now posted on various sites of the internet. One such site has the title "*Shh* Mouse Gene Detail - MGI:98297 sonic hedgehog". Incidentally, the acronym MGI stands for Mouse Genome Informatics, which is a data base maintained by The Jackson Laboratory of Bar Harbor, Maine. When this phrase is entered into a search engine for the internet, an outline of various subtopics appears. Included in these subtopics are links to data bases for gene models. One such model had the title VEGA Gene Model with instructions to download in various formats, such as ADOBE pdfs, if an investigator were interested in more detailed information. The sequence information for this gene was downloaded and it was found that the *Shh* genomic region was made of a little more than 11,461 nucleotides that were listed. Within this region there are three exons of varying lengths, which were displayed in red letters and are known to be coding regions. Evidently, within the regions it seems probable that transcription products may involve alternate splicing. Also contained in this region were 4 introns and it appeared that a majority of the nucleotides in this region were included in the introns. From this information, one may conclude that to the 1 Mb one should add about 11,461 nucleotides for a model of the *Shh* genic region involved in coding for proteins.

On the *MGI* web site there is also a page giving details regarding the *Lmbr*1 genomic region on chromosome 5 and from this page it was also possible to download the sequenced version of this region, which contained a little more than 147,361 nucleotides. Among these nucleotides, 18 exons were indicated and many of these exons consisted of relatively few nucleotides. For the most part, by nucleotide count, the genomic region *Lmbr*1 contains mostly introns and, evidently, within one of these intronic regions the regulatory region $MFCS1$ occurs. The approximate size of this region is 1167 *bp*, see Sagai *et al.* (2005) for details. Evidently, some of the 1 Mb separating the *Shh* genome region with its enhancer $MFCS1$ includes many of the nucleotides in the genomic region *Lmbr*1 so that the number of nucleotides that must be included in a model of this system is of order $1\ Mb + 11,461$ bases. Therefore, in order to simulate the process of nucleotide substitution and other types of mutation, a computer with several giga bytes of memory would be an adequate platform on which to conduct computer simulation experiments.

One of the important problems in the discovery of regulatory regions among millions of bases of a genomic region is to induce mutation in noncoding regions and then observe whether these mutations are implicated in the development of some part of an embryo. An interesting paper in which ideas of this kind were implemented was that Masuya *et al.* (2007), who induced mutations in the regulatory region $MFCS1$ of the mouse genome. An extensively used method for inducing mutations is known by the acronym ENU, which stands for the chemical N-eythyl-N-nitros urea that can induce point mutations or single nucleotide substitutions, see the review of Justice *et al.* (1999) for details. This was the method used by Masuya *et al.* (2007) in their work to induce four new single nucleotide substitutions. Their results suggested that mutations symbolized by $M101116$ and $M100081$ affect the regulatory activity of the region $MFCS1$, which suppresses anterior *Shh* expression in developing limb buds. This study was one of many that demonstrated that ENU induced mutations in noncoding regions was an effective approach for exploring the function of conserved noncoding regions of genome. To find more references on the use of this procedure to induce mutations, it is suggested that an interested reader enter the phrase "ENU Mutagenesis" into a search engine on the internet.

Among the many topics listed for each gene in the *MGI* database is that on mammalian homology, which is a set of concepts relating to the idea that similar genomic structures in different species may be interpreted as being evolved from a common ancestor. For the case of the *Shh* gene, the species listed under mammalian homolog are humans, chimpanzees, cattle, domestic dogs and rats. These species are also mentioned for other mouse genes described in the *MGI* data base. Another link on the page of many genes is the term mammalian orthology. Orthologs and paralogs are two types of homologous sequences. Orthology is devoted to the description of genes in different species that are thought to have been derived from a common species. Orthologous genes may not have the same function. Paralogy describes genes within a single species that have diverged by gene duplication. Further information on these concepts is available on the internet. These concepts help to justify using species such as mice, which may be manipulated experimentally, to get some ideas of about the function homologous genes in humans because many types of experimentation with humans are banned for moral reasons.

An interesting example of homology is the Sonic Hedgehog gene in humans, which is denoted by *SHH* and has been implicated in holoprosencephaly, see Nanni *et al.* (1999) for details. Holoprosencephaly, *HPE*, is a common developmental anomaly of the human forebrain and midface, where the cerebral hemispheres fail to separate into distinct left and right halves. The numerous authors of this paper performed a mutational analysis of the complete coding region and intron-exon junction of the *SHH* gene in 344 unrelated individuals as well as an additional 13 unrelated individuals with *SHH* mutations. Included in these mutations were nonsense and missense mutations, deletions and insertions, which were distributed throughout the genomic region *SHH*. A summary of the mutations observed in several studies by the authors were listed in Table 1 of the paper. Apart from an insertion and several deletions, the mutations listed were nucleotide substitutions in three letter codons, which resulted in different amino acids being inserted into a protein chain. There were also mutations that led to stop coding signals. Among other things, the authors observed that mutations in the *SHH* genomic region seem to be the cause of a significant proportion of autosomal dominant holoprosencephaly. Several pictures of infants with this anomaly were displayed in the paper and the authors suggested that the final phenotypic expression of the gene for a given individual may depend on the interactions of several gene products as well as environmental factors.

A related and interesting paper on the regulation of a remote Sonic hedgehog forebrain enhancer was the paper by Jeong *et al.* (2008). It is known that the secreted morphogen, Sonic hedgehog, *Shh*, is a significant determinant of brain size and craniofacial morphology. In humans *SHH* haploinsufficiency results in holoprosencephaly, HPE, which is a defect in anterior midline formation. Even though the importance of maintaining SHH transcript levels above a critical level is known, little is known about the upstream regulators of SHH expression in the forebrain. These authors used a combination of genetic and biochemical experiments to uncover a critical pair of cis and trans acting determinants of Shh forebrain expression. As it turned out a rare nucleotide variant located 460 kb upstream of SHH was discovered in an individual with HPE that resulted in the loss of Shh brain enhancer-2, $SBE2$, activity in the hypothalamus of transgenic mouse embryos. The $SBE2$ sequence was screened for DNA binding proteins, which led to providing a direct link between Si$x3$, a homeoprotein, and Shh regulation during forebrain development and in the pathogenesis of HPE. Further technical details may be obtained from the paper by Jeong *et al.* (2008). The results of this paper also demonstrate that the homologies between mouse and human genes can lead to insightful results. Also illustrated in the paper are the difficulties that are often encountered when attempts are made to find regions that are binding sites for regulatory proteins and related chemical structures.

Bibliography

[1] Akiva, P., Toporik, A. *et al.* (2006) Transcription-mediated gene fusion in the human genome. Genome Res. **16**:30–36.

[2] Amano, T. *et al.* (2007) Chromosomal dynamics at the *Shh* locus:Limb bud-specific differential regulation of com(2009) petence and active transcription. Developmental Cell **16**:47–57.

[3] Barbu, S. and Limnois, N. (2008). **Semi-Markov Chains and Hidden Semi-Markov Models Towards Applications - Their Use in Reliability and DNA Analysis**. Lecture Notes in Statistics, Springer.

[4] Beadle, G. W. and Tatum, E. L. (1941). Genetic control of biochemical reactions in Neurospora. Proc. Nat. Acad. Sci. **27**:499–506.

[5] Benzer, S. (1955) Fine structure of a genetic region in bacteriophage. Proc. Nat. Acad. Sci. **41**:344–354.

[6] Berget, S. M., Moore, C. and Sharp, P. A. (1977) Spliced segments at the $5'$ terminus of adenovirus 2 late $mRNA$. Proc. Natl. Acad. Sci. **74**:3171–3175.

[7] Bumenthal, T. (2005) Trans-splicing and operons. WormBook (ed. The C. elegans Research Community). WormBook, doi/10.1895/wormbook.1.5.1, http://www.wormbook.org.

[8] Boys, R. J. *et al.* (2000) Detecting homogeneous segments in DNA sequences by using hidden Markov models. Appl. Statist. **49**:269–285.

[9] Boys, R. J. and Henderson, D. A. (2002) On determining the order of Markov dependence of an observed process governed by a hidden Markov model. Scientific Programming **10**:241–251.

[10] Boys, R. J. and Henderson, D. A. (2004) A Bayesian Approach to DNA Sequence Segmentation. Biometrics **60**:573–588.

[11] Carroll, S. B. *et al.* (2008) Regulating Evolution. Scientific American, May: 61–67.

[12] Cech, T. R. (2004) Exploring the New RNA World. Nobelprize.org

[13] Chow, L. T. *et al.* (1977) An amazing sequence arrangement at the $5^{'}$ ends of adenovirus 2 messenger RNA. Cell **12**:108.

[14] Crick, F. H. C. (1958) On protein synthesis. Symp. Soc. Exp. Biol. XII:138–163.

[15] Doolittle, R. (1986) **Of URFs and ORFs: A primer on how to analyze derived amino acid sequences**. University Science Books, Mill Valley, CA.

[16] Du, J. *et al.* (2006) A supervised hidden Markov model framework for efficiently segmenting tiling array data in transcriptional chIP-chip experiments: systematically incorporating biological knowledge. Bioinformatics **22**:3016–3024.

[17] **The ENCODE Project: ENClycopedia Of DNA Elements**. http://www.genome.gov/1005107.

[18] http://web.archive.org/web/20050428090317/www.ensembl.org/Genesweep

[19] Fiers, W. *et al.* (1971) Recent progress in the sequence determination of bacteriophage MS2 RNA. Biochemie **53**:495–506.

[20] Fiers, W. *et al.* (1976) Complete nucleotide sequence of bacteriophage MS2 RNA: Primary and secondary structure of the replicase gene. Nature **260**:500–507.

[21] Fleischmann, R. D., Adams, M. D. *et al.* (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd., Science **269**:496–512.

[22] Gelinas, R. E. and Roberts, R. J. (1977) One predominant $5^{'}$-undecanucleotide in adenovirus 2 late messenger $RNAs$. Cell **11**:533–544.

[23] Gerstein, M. *et al.* (2007) What is a gene, post-ENCODE? History and undated definition. Genome Research **17**:669–681.

[24] Griffith, F. (1928) The significance of pneumococcal types. J. Hyg. (Lond.) **27**:113–159.

[25] Griffiths-Stotz (2006) Genes in the post-genomic era. Theor. Med. Bioieth. **27**:499–521.

[26] Heber, J. *et al.* (2002) Splicing graphs and EST assembly problem. Bioinformatics **18**:S181–S188.

[27] Hershey, A. D. and Chase, M. (1955) An upper limit to the protein content of the germinal substance of bacteriophage T2. Virology: **1**:108–127.

[28] Jacob, F. and Monod, J. (1961) Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol. **3**:318–356.

[29] Jensen, S. T. *et al.* (2004) Computational discovery of gene regulatory binding motifs: A Bayesian Perspective. Statistical Science **19**:188–204.

[30] Jeong, Y. *et al.* (2008) Regulation of remote Sonic hedgehog forebrain enhancer by Six3 homeoprotein. Nature Genetics **40**:1348–1353.

[31] Justice, M. K. *et al.* (1999) Mouse *ENU* mutagenesis. Human Molecular Genetics **8**:1955–1963.

[32] Kapranov, P., Cawley, S. E. *et al.* (2002) Large-scale transcriptional activity in chromosomes 21 and 22. Science **296**:916–919.

[33] Lang, X. Y., Wang, J. and Chi, X. B. (2008) The research progress of tiling array technology and applications. Chinese Science Bulletin **53**:817–824.

[34] Lander, E. S., Linton, L. M. *et al.* (2001) Initial sequencing and analysis of human genome. Nature **409**:860–921.

[35] Li, W. *et al.* (2005) A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and it application to p53 sequences. Bioinformatics **21**:i274–i282.

[36] Lindblad-Toh, K. C. M., Wade, T. S. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature: **438**:803–819.

[37] Lodish, H., Scott, M. P. *et al.* (2000) **Molecular Cell Biology, Fifth Edition**. Freeman and Co., New York.

[38] Marioni, J. C. *et al.* (2006) BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. Bioinformatics **22**:1144–1146.

[39] Masuya, H. *et al.* (2007) A series of *ENU*-induced single base substitutions in a long-range cis-element altering Sonic hedgehog expression in the developing mouse limb bud. Genomics **89**:207–214.

[40] McClintock, B. (1929) A cytological and genetical study of triploid maize. Genetics **14**:180–222.

[41] McLachlan, G. J., Do, K. A. and Ambroise, C. (2004). **Analyzing Microarray Gene Expression Data**. John Wiley & Sons, Inc., Hoboken, N. J.

[42] Morgan, T. H., Sturtevant, A. H. Muller, H. J. and Bridges, C. B. (1915) **The Mechanism of Mendelian Heredity**. Holt Rinehart & Winston, New York.

[43] Mount, D. W. (2001) **Bioinformatics - Sequence and Genome Analysis**. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

[44] Muller, H. J. (1927) Artificial transformation of the gene. Science **46**:84–87.

[45] Nanni, L. *et al.* (1999) The mutational spectrum of Sonic Hedgehog gene in holoprosencephaly: *SHH* mutations cause a significant proportion of autosomal dominant holoprosencephaly. Human Molecular Genetics **8**:2479–2488.

[46] Nirenberg, M., Leder, P., Bernfield, M., Brimacome, R., Trupin, J., Rottman, F. and O'Neal, C. (1965) *RNA* code words and protein synthesis, VII, On the general nature of *RNA* code. Proc. Natl. Sci. **53**:1161–1168.

[47] Ohno, S. (1972) So much "junk" *DNA* in the genome. In Evolution of genetic systems. vol. 23 (ed. H. H. Smith), Brookhaven Symposium in Biology. Gordon and Breach, New York.

[48] Parra, G., Raymond, A. *et al.* (2006) Tandem chimerism as a means to increase protein complexity in the human genome. Genome Res. **16**:37–44.

[49] Pearson, H. (2006) Genetics: What is a gene? Nature **441**:398–401.

[50] Ponjavic, J. P., Pointing, C. P. and Lunter, G. (2007) Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. Genome Res. **17**:556–565.

[51] Prud'homme, B. *et al.* (2007) Emerging Principles of Regulatory Evolution. PNAS 104, Supple **1**:8605–8612.

[52] Pulst, S. M. (1999) Genetic linkage analysis. Arch Neuol. **56**:667-672.

[53] Rinn, J. L., Enskirchen, G. *et al.* (2003) The transcriptional activity of human chromosome 22. Genes & Dev. **17**:529–540.

[54] Rogic, S. *et al.* (2001) Evaluation of gene finding programs on mammalian sequences. Genome Res. **11**:817-832.

[55] Roll-Hansen, N. (1989) The crucial experiment of Wilhelm Johannsen. Biol. Phlois **4**:303–329.

[56] Sagai, T. *et al.* (2005) Elimination of long-range cis-regulatory module causes complete loss of the limb-specific *Shh* expression and truncation of the mouse limb. Development **4**:797–803.

[57] Searls, D. B. (1997) Abstract: Linguistic approaches to biological sequences. Comput. Appl. Biosci. **13**:333–344.

[58] Searls, D. B. (2001) Reading the book of life. Bioinformatics **17**:579–580.

[59] Searls, D. B. (2002) The language of genes. Nature **420**:211–217.

[60] Sinnott, E. W., Dunn, L. C. and Dobzhansky, Th. (1950) **Principles of Genetics**, McGraw Hill, New York, Toronto and London.

[61] Snustad, D. P. and Simmons, M. J. (2006) **Principles of Genetics - Fourth Edition**. John Wiley and Sons, Inc.

[62] Soll, D. *et al.* (1965) Studies on polynucleotides, XLIX. Stimulation of the binding of aminoacl-sRNAs to ribosomes by ribonucleotides and a survey of codon assignments for 20 amino acids. Proc. Natl. Acad. Sci. **54**:1378–1385.

[63] **Transcription:** *BiologyPages/T/Transcription.html*

[64] **Translation:** *BiologyPages/T/Translation.html*

[65] Tournamille, C. *et al.* (1995) Disruption of the *GATA* motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. Nat. Genet. **10**:224–228.

[66] Venter, J. C., Adams, M. D. *et al.* (2001) The sequence of the human genome. Science **291**:1304–1351.

[67] Wain, H. M., Bruford, E. A. *et al.* (2002) Guidelines for human gene nomenclature. Genomics **79**:464–470.

[68] Waterston, R. H., Lindblad-Toh, K. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. Nature **420**:520–562.

[69] Watson, J. D. and Crick, F. H. C. (1953) A structure of deoxyribonucleic acid. Nature **171**:964–967.

Chapter 14

Detecting Genomic Signals of Selection and the Development of Models for Simulating the Evolution of Genomes

14.1 Introduction

This chapter has been assembled around several but related themes, regarding the evolution of the genome of a given species. Included in these themes are definitions of what is meant by the detection of signals of positive or negative selection and the meaning of the terms positive or negative selection as well as the developments statistical methods for the detection of signals of natural or artificial selection. Attention will also be given to a brief review of the development of algorithms to simulate the evolution of a model genome. These algorithms are based on two approaches to viewing evolution in time. One approach looks backward in time starting from the present to some time in the distant past and is usually referred to under the rubric of coalescence. A second approach starts at some time point in the past and projects forward in time to some more recent time in the past or the present, which is the approach that has been followed in the preceding chapters of this book. Due to the rather large array of related topics that were included in this chapter, it contains more pages than any other in this book and is thus a booklet within a book. Nevertheless it was deemed advisable to present all these topics as a whole in one chapter rather than splitting them into several short chapters.

The recent research projects reviewed in this chapter on the detection of signals of positive or negative selection were based on two types of data that were available in the public domain. The first type considered was haplotype data, which consists of relatively short segments of human DNA that were used to detect signals of positive or negative selection. A second type of data included sequenced genomes of several mammalian species, including man and chimpanzees. The analysis of human haploid data centered around the development of numerical scores, which were thought to provide some measures of the action of selection, and in this connection, the authors also used estimates of probability density function that was estimated by using a coalescence model which simulated the evolution of genomic sequence back in time under the influence of natural selection on a time scale of generations. From the point of view of the development of statistical methods, the estimation and application of this density to detect signals of selection was perhaps the most novel part of the research based on haploid data. In subsequent sections of this chapter the technical details involved in this research will be discussed in greater detail.

Unlike the analysis of haploid data, which was carried out on a time scale of generations, the detections of signals of natural or artificial selection within the sequenced genomes of several mammalian species were based on a evolutionary time scale expressed in continuous time. Just as in chapter 6, it was assumed that nucleotide substitutions or the evolution of codons in the protein coding genes under consideration evolved in continuous time according to a Markov jump process with a state space consisting of types of three letter codons. Given this Markov process, it was possible to derive likelihood functions of the data and estimate the parameters of the model, using the method of maximum likelihood. Scores based on maximum likelihood estimates were then used to detect signals of selection as well a phylogenetic relationships among the species, which will be discussed in more detail in subsequent sections of this chapter. By any measure, the scope of the phylogenetic study based on sequenced data of the genomes of several species was greater than the study involving haploid data discussed above.

As the papers on the models dealing simulating the evolution of model genomes along with the supporting documents were reviewed, it was concluded that the documentation of the algorithms used in these software packages was inadequate from the point of view of developing Monte Carlo simulation model with concepts anchored in modern theories of stochastic process and statistical methods for informative presentations of the results of Monte Carlo simulation experiments. Consequently, a decision was made to devote several sections of the chapter to the description and development of a set of preliminary algorithms that would form the basis for the development of software to simulate the evolution of a model genome on a time scale of generations. For diploid populations, the focus of attention in these algorithms is the biological process of meiosis, which gives rise to the types of gametes produced by an individual. In one phase of meiosis, the DNA content of a cell is doubled and it is during this phase that there are risks that several types of mutations may occur. Included in these types of mutations were nucleotide substitutions, deletions, insertions, duplications and inversions. Following the doubling of the content of DNA in a cell, pairs of chromosomes align on a spindle and at this point genetic recombination among maternal and paternal DNA may occur. Two types of genetic recombination were included in these algorithms; namely reciprocal crossing over and non-reciprocal crossing over, which is also known as gene conversion. Given genetic recombination, there are conditional probabilistic algorithms describing the production of gametes for reciprocal and non-reciprocal crossing over. In non-diploid species, which reproduce asexually by mitotic cell divisions so that genetic recombination does not occur, the algorithms describing the various types of mutation could also be utilized in simulating the evolution of their genomes.

The development of methods describing the evolution of genomes through mutations and rearrangements of DNA seems to be a relatively new and emerging field. An interesting paper that provides a review of recent papers on the simulation of genomic evolution is that of Fletcher and Yang (2009). Another set of references on stochastic models of genomic rearrangements and other mutations in genomes may be found in the book Durrett (2008) in chapter 9.

14.2 Types of Selection and Genomic Signals

In chapters 4 and 5, which were devoted to Wright-Fisher processes characterized by constant population size from generation to generation, natural selection was defined in terms of probabilities that alleles or gametes would be contributed by a parental generation to the next generation of offspring. In chapters 10,11 and 12, however, components of natural selection were defined in terms of parameters of multitype self regulating branching process, and in computer experiments, based on these models, mutation and selection were quantified by assigning numerical values to parameters. Among these parameters for non-age dependent branching processes were the expected number of offspring produced by each type of individual per generation and a parameter characterizing an individual's ability to compete in terms of population density. When two sexes were accommodated in a formulation, such as those in chapters 11 and 12, another component of natural selection included in the formulation was a module describing the selection of mates. Selection of mates in turn could depend on the phenotypes of the females and males involved in the mating process. In either stochastic or deterministic computer simulation experiments based on these models, the aim of these experiments was to see if some type rose to high prevalence in a population or if several of types were maintained in a population after a large number of generations.

However, when attempting to search the human genome and that other species for signatures of either natural or artificial selection, the idea of selection is usually defined in abstract operational terms rather than in terms of the parameters stochastic evolutionary models such as either the Wright-Fisher process, the class of branching processes under consideration or even deterministic evolutionary models. It is interesting to note however, as will become apparent in what follows, that the rationale underlying tests for selection seems to have been influenced by experimental results obtained from computer experiments using mathematical evolutionary models. According to these operational definitions, positive natural or artificial selection is defined as the force that drives the increase in prevalence of some advantageous trait. The nature of this force is clear in artificial selection, when desirable traits in plants or animals, determine which parents that will be selected by breeders to produce the next generation, but, in natural populations, the nature of this force is less clear even though it is possible to get a partial handle on it in terms of parameters governing a stochastic evolutionary process.

For the case of diploid populations such as humans, selection is referred to as balancing in those case in which some number of alleles or genotypes attributable to some locus are present in a population. Other terms are also used to describe natural selection. One of these is negative selection, which refers to selective removal of alleles that are deleterious, which are thought to arise through the process of mutation. Such selection is also referred to as purifying selection. Contained in the online supporting material of the paper by Sabeti *et al.* (2006) is table S1 in which 91 genes or genomic regions of the human genome are listed as putative sites where natural selection has acted in the past or is acting in the present. Also contained in this table is a column giving the chromosomal location of each of the 91 genomic regions.

Another component of this table is the idea of a signal that is interpreted as a manifestation of natural selection in some genomic region. Included in the idea of signals of natural selection are such terms as long haplotypes, diversity, function-altering mutations, derived allele frequency and population differentiation. The idea of a long haplotype and its ramifications for natural selection have been discussed in the letter to Nature by Sabeti *et al.* (2002). The first step in identifying such objects as long haplotypes is to select some locus of interest. The next step is to assess the age of each core haplotype by the decay of its association with alleles at various distances from the locus, as measured by extended haplotype homozygosity (EHH). Those haplotypes that have unusually high EHH are indicators of mutations that rose to prominence in the human gene pool faster than that expected under neutral evolution.

An example of a genomic region with a long haplotype signal is that of gene 1 with the symbol $ABCC1$ on chromosome 16, see table S1 of Sabeti, Schaffner *et al.* (2006). A function associated with this gene or region is that of one of a family or organic anion transporters, which can protect cells against drugs. In mice these transporters have been linked to Streptococcus pneumonia resistance. Listed in the column of table S1 for this type of natural selection for this genomic region was the term positive. There is also a column for tests, statistical and otherwise, used to detect selection, which included the symbols, F_{ST}, LRH and MAF threshold. The symbol F_{ST} stands for a statistical function of data on haplotype and other frequencies such as genes in a population, which will be discussed in a subsequent section. The symbol LRH stands of luteinizing hormone releasing hormone and the symbol MAF stands for minor allele frequency. If a reader wishes to obtain more information on these symbols, they may be entered into a search engine for the internet. The last column of the table S1 lists the reference or references in which evidence for the type natural selection indicated was reported.

The genomic region or gene labeled 2 in the Sabeti *et al.* (2006) paper was the ABO blood type region on chromosome 9, which was discussed in some detail in chapter 13 so that no further details as to the function of these alleles will be given here. For this region the signal for natural selection was listed as diversity (high), and the type of selection listed was designated as balancing. Evidently, the rationale for using the term balancing selection is that in human populations around the globe the frequencies of the A, B and O alleles seem to be relatively constant within some geographic region but may vary among regions. Among the several types of statistical test listed for detecting this signal of natural selection was F_{ST} and others that will be discussed in a subsequent section.

Gene or genomic region listed as 5 in the Sabeti *et al.* paper was designated by the symbol $ADAM2$ and is located on the X chromosome. The signal for natural selection acting on this region was listed as function altering mutations, and the type of natural selection acting at this locus was listed as positive, purifying. Of the two types of tests listed for detecting this type of selection was the acronym $PAML$, which according to an internet search, stands for phylogenetic analysis by maximum likelihood. Briefly, this is a type of statistical procedure was designed to estimate a most likely phylogenetic tree at the molecular level associated with the genomic region under consideration. Evidently, this acronym also stands of a software package produced by Ziheng Yang, which was described among the internet sites that appear when the symbol $PAML$ is typed into a internet search engine. The function of this gene or genomic region is thought to be that of coding for heterodimeric sperm protein, which plays a potential role in the fertilization of an egg and thus has implications for reproductive success of a male. Function altering mutations may include nucleotide substitutions which change an amino acid three letter code or, perhaps, deletions and the resulting frameshifts may be the mechanisms underlying the mutations in question. A reader may wish to consult the paper cited in connection with the genomic region $ADAM2$ for specific details concerning the mutations under consideration.

The genomic region or gene on human chromosome 1 was number 7 in table S1 of Sabeti *et al.* (2006) and had the label AGT. In this case the signal for natural selection was derived allele frequency and the gene function was salt regulation pathway associated with hypertension. The type of selection was termed positive and several types of statistical tests were listed in connection with testing for natural selection as this genomic region. To get some idea of the meaning of the term derived allele frequency, it is recommended that the paper of Fry *et al.* (2006) be consulted. In this paper the authors used haplotype-based techniques to estimate the relative age of alleles in connection for screening for signals of recent positive selection. By using simulations and empirical data from the International HapMap Project, these authors showed that a simple pair-wise metric of haplotype homozygosity gives significantly higher mean values for human single-nucleotide-polymorphisms alleles that appear to be have been derived from those thought to be ancestral when compared with the chimpanzee genome.

The last illustrative example taken from table $S1$ of Sabeti *et al.* (2006) is that for genomic region or gene numbered 8 with the label $ALDH2$

on human chromosome 12. For this genomic region the signal of natural selection was listed as population differentiation and its function was alcohol metabolism, with associated alcoholism, and the type of selection was listed as positive. The statistical tests used to detect evidence for positive selection were listed as F_{ST} and LD, where LD stands for linkage disequilibrium. Previously, in chapter 3, convergence to linkage equilibrium under the assumption of no mutation or selection was studied for case of two to many liked loci in large populations. Consequently, the idea underlying such tests is that if one can demonstrate that neighboring loci on the same chromosome are in linkage disequilibrium, this would be putative signal for the action of mutation and or selection. These ideas will be discussed further in a subsequent section of this chapter.

In concluding this section, it will be instructive to suggest some ways in which the stochastic evolutionary models presented in chapter 9 through 12 would be useful in the further assessing levels of uncertainties that arise in research designed to detect signatures of natural selection in the human and other genomes. As suggested in one of the examples discussed above involving long haplotypes, one approach to detecting natural selection is to conceptually compare a population undergoing mutations in which it was thought that natural selection had or was acting with that which would happen in a population subject to mutation but no selection. Such ideas could easily be quantified in terms of assigning plausible values to parameters in the multitype stochastic models described in chapters 10, 11 and 12.

Initially, for the sake of simplicity, the multitype models presented in chapter 10 could be used to design computer experiments applicable to evolution of populations of bacteria, for example, in which in one set of experiments parameter values were chosen such that mutation occurred but there was no selection. The results of these experiments could then be compared to those in which parameter values were chosen such that mutations and selection were acting simultaneously. In preliminary experiments, the nonlinear models embedded in stochastic processes could be used to provide informative overviews of the evolution of population under the two sets of assumptions. Then, Monte Carlo simulation experiments could be run in confirmatory experiments designed to take into account levels of stochasticity that are thought to be central to evolution of biological populations and then compare statistical summaries of the Monte Carlo simulation data with the trajectories computed, using the embedded deterministic models.

Such comparative experiments could also be done, using the two sex models discussed in chapters 11 and 12, but the technical problems that arise in summarizing data generated in such experiments, particularly when Monte Carlo simulation date generated from the age-dependent models in chapter 12 is considered, would be greater than those based on the simpler models discussed in chapter 10. One of the difficult challenges in the further development multitype models, would be that of extending the mutations described at the Mendelian level in chapters 10,11 and 12 to the many types of mutations that can be observed at the molecular level when DNA is sequenced. Some preliminary ideas involving generational time scales and the mutational process of nucleotide substitution will be discussed in a subsequent section of this chapter.

Another type of computer simulation experiment that could be done based on the models described in chapter 10,11 and 12 is that of simulating the age distribution of a mutation in a simulated population. Evidently, among the types of mutation simulated could be those in long haplotypes. Briefly, in such an experiments the ideas discussed in chapter 9 on coalescence with only one type of individual under consideration would need to be extended to the cases of multitype processes with mutation. By simulating realizations of these multitype branching processes, one could represent them in a computer as genealogies extending over several thousand generations and emanating from a single individual, couple or a small founder population. By choosing a sample at random of size $N \geq 1$ individuals carrying this mutation in some generation called the present, one could program the computer to follow the linage of each individual in the simulated genealogy to that ancestor in which the mutation first occurred and was passed on to at least one of his offspring. It would, of course, also of interest to search the simulated genealogy for evidence that the mutation under consideration could have also occurred independently in at least two individuals. It is suggested to a reader or group, who may wish to work through such a project, that the developmental work begin with the more simpler class of multitype branching processes described in chapter 10.

14.3 An Overview of Models for DNA Sequence Evolution in Large Genomic Regions

In the preceding section, an overview of genomic signals that were used to detect the action of natural selection, particularly in human populations, was given along with some suggested examples as to how working within a branching process paradigm may yield interesting answers to questions that arise when contemplating the evolution of populations under going mutation and selection. One of the suggested examples was that of estimating the distribution of the age of a mutation by Monte Carlo simulation methods by using one or more classes of multitype branching processes. It was also mentioned that one of the challenges encountered when working within the branching process paradigm was that of incorporating into a formulation a large genomic sequence so that the effects of mutation and selection could be studied at the molecular level. In this section, a brief review of the existing literature will be given on the subject of doing population simulations over large genomic regions and applying them to such quests as finding signatures of natural section in the human genome.

Hoggart *et al.* (2007) have emphasized that computer simulation is an invaluable tool for investigating the effects of various assumptions used in population genetic models which have been proposed as mechanisms to explain observed diversity in a genomic region under consideration. Simply stated, given a set of observations on a genomic region sampled from an existing population, the problem is to look backward in evolutionary time to get some idea of events that may have occurred in the past to provide an explanatory basis for the observations being made in the present. In terms of generations, back in time could mean thousands of generations. From the point of view of constructing computer simulation models, there are at least two approaches. One approach looks from the present back in time to an ancestor in which the event, usually some type of mutation, occurred, which is the approach that was reviewed briefly in chapter 9 and referred to as coalescence. Theoretical developments of these ideas have led to the development of several software packages, which have been widely used. Among these software packages are *MS* (Hudson 2002), *SELSIM* (Spencer and Coop (2004)) *Coa* Si *m* (Mailund *et al.* (2005)) and *FastCoal* (Marjoram and Wall (2006)). However, Hoggart *et al.* (2007) and others have argued that coalescent methods have significant limitations, and, among these limitations, is that of including genetic recombination in a formulation, particularly when genomic regions with a large number of bases are being considered.

An alternative to backward in time methods is forward in time methods, which has been the main theme in the preceding chapters of this book. Others who have worked from the point of view of this perspective include Peng and Kimmel (2005) and Peng *et al.* (2007). Another group of authors working within the forward in time perspective are those in Hoggart *et al.* (2007), who have used the acronym *FREGENE* as a name for their software package. Unlike some coalescent approaches, $FREGENE$ is not limited gamete sampling models at a single locus, constant population size or small genomic regions, but has the capability of accommodating diploid inheritance with linkage and genetic recombination. Included in the term "genetic recombination" are both the processes of crossing over and gene conversion. Unfortunately however, in a document devoted to the technical aspects underlying their software package, most attention is directed toward technical details regarding the use of the software and not to the mathematical ideas under pinning the software written for the computer simulation of such processes as genetic recombination and mutation. It should also be mentioned that the software package, $FREGENE$, appears to limited to a deterministic paradigm rather than with the classes of stochastic processes with their embedded deterministic models, which have been considered in chapters 9,10,11 and 12 of this book.

These authors state that included in their term "genetic recombination" is the process of gene conversion. It, therefore, necessary to include an overview of the process of gene conversion so that this biological phenomena my included in the mathematical description of the process of genetic recombination. Consider, for example, a diploid individual with two parents. Suppose the female contribution to a region of a genome of this individual may be represented in the form of two chromatids

$$\begin{array}{|c|c|c|c|c|}\hline 5' & D & E & F & 3' \\ \hline 3' & D' & E' & F' & 5' \\ \hline \end{array}\,, \tag{14.3.1}$$

and suppose the male contribution to the genomic region of this individual may be represented as the following two chromatids

$$\begin{array}{|c|c|c|c|c|}\hline 3' & d' & e' & f' & 5' \\ \hline 5' & d & e & f & 3' \\ \hline \end{array}\,. \tag{14.3.2}$$

As indicated in the symbolism, it is supposed that each pair of chromatids is represented at the DNA level and the symbols D, E, F and d, e, f with or without primes represent strands of DNA. The number of bases in each of these strands may vary, but it will be assumed that homologous strands have the same number of bases. The genotype of the individual under consideration has developed from the gametes of the parents represented in (14.3.1) and (14.3.2). During meiosis this individual will produce gametes containing various combinations of the maternal and paternal DNA.

An illustrative symbolic example of gene conversion, which may occur during meiosis, is displayed in the following 4×5 table. Each row of this table is interpreted as a type of gamete that may be produced by the individual under consideration.

$5'$	D	e	F	$3'$
$3'$	D'	e'	F'	$5'$
$3'$	d'	e'	f'	$5'$
$5'$	d	e	f	$3'$

(14.3.3)

Observe that the DNA segments e' and e, which originated in the male gamete displayed in (14.3.2), have replaced the DNA segments E and E' that were present in the female gamete (14.3.1). It is interesting to note when the process of gene conversion occurs all gametes carry only the DNA segments e' and e, which, in this example, were contributed by the male parent.

When the process of meiosis is normal, that is gene conversion does not occur, then by the process of crossing over the gametes of an individual may contain various combinations of the maternal and paternal DNA. An illustrative symbolic example of crossing over occurring during meiosis is displayed in the 4 × 5 table.

$5'$	D	E	f	$3'$
$3'$	D'	E'	f'	$5'$
$3'$	d'	e'	F'	$5'$
$5'$	d	e	F	$3'$

(14.3.4)

Just as in (14.3.3), each row of this table will be interpreted as a type of gamete that may be produced by the individual under consideration. The process of crossing over is reciprocal in the sense the DNA segments f and f' contributed by the male parent have been inserted into the DNA contributed by the female parent as is illustrated in the two upper rows of the table. Similarly, the two DNA segments F and F', which were originally part of the DNA contributed by the female parent, have been inserted into the DNA originally contributed by the paternal parent as illustrated in the lower two rows of table (14.3.4). The reader will remember that in chapter 2, the process of reciprocal genetic recombination has been described in detail, in terms of a mathematical formulation with respect to many loci and with many alleles at each locus.

Another illustrative example depicting the process of gene conversion may be demonstrated by considering the genotype symbolized in the 4 × 5 table

$5'$	D	E	F	$3'$
$3'$	D'	e'	F'	$5'$
$3'$	d'	E'	f'	$5'$
$5'$	d	e	f	$3'$

. (14.3.5)

Just as before, the upper two rows of the table will be interpreted as the DNA contributed by the female parent and the lower two row the DNA contributed by the male parent. The type of configuration presented in (14.3.5) is sometimes referred to collectively as heteroduplexes.

For this case, the process of gene conversion occurring is symbolized by the types of gametes in the rows of the following table.

$5'$	D	e	F	$3'$
$3'$	D'	e'	F'	$5'$
$3'$	d'	e'	f'	$5'$
$5'$	d	e	f	$3'$

(14.3.6)

From this table, it can be seen that the DNA segment e, which was originally part of the DNA contributed by the male to the genotype represented in (14.3.5), replaced the DNA segment E in the top row of the table that was part of the DNA contributed by the female parent of the individual represented in (14.3.5). Similarly, the DNA strand E' which was part of the paternal contribution to the individual represented in (14.3.5) was replaced by the DNA strand e' in (14.3.6), which was part of the material contribution to the DNA of the individual represented in (14.3.5). At this point in the discussion, it is of interest to note that if both the strands of DNA e and e' contained a recessive allele a, then all gametes represented in (14.3.6) would contain only the allele a. Similarly, suppose both strands of DNA E and E' carry an allele A which is dominant to a. Then, all offspring arising from the union of gametes of the type displayed in (14.3.6), would express on the recessive phenotype determined by the genotype aa. This illustrative example demonstrates how the process of gene conversion can distort expected Mendelian ratios in crosses for which the genotype of each parent is known.

The process of gene conversion is thought to occur when the enzymes involved in editing newly copied DNA do not function properly. A case in which the editing enzymes were able to repair the gametes generated by an individual with the genotype symbolized in (14.3.5) is presented in the following table in which it is indicated that the process of gene conversion did not occur.

$$\begin{array}{|c|c|c|c|c|}\hline 5' & D & E & F & 3' \\ \hline 3' & D' & E' & F' & 5' \\ \hline 3' & d' & e' & f' & 5' \\ \hline 5' & d & e & f & 3' \\ \hline \end{array} \tag{14.3.7}$$

If it is again supposed that DNA strands E and E' carry a dominant allele A and the strands e and e' carry the recessive allele a, then when a large number of gametes of the form displayed in the rows of (14.3.7) are combined to produce offspring, one would expect to observe the standard Mendelian ratio 3 : 1 of dominant phenotype $A-$ to the recessive phenotype aa in a large sample of offspring arising from the union of gametes depicted in (14.3.7). The material on gene conversion presented in this section was adapted from the web site $web-books.comCh8D4$. In a subsequent section, some preliminary ideas on the stochastic modelling of the process of gene conversion will be presented.

The interpretation of present day observations on a region DNA in terms of evolutionary events that happened in the distant past is closely intertwined with existing theories of recent human evolution and that of other species. Theories as to the recent natural history of our species, *Homo sapiens*, has been put forth by a number of authors working in genetics and other fields such as anthropology and paleontology. There is a vast and growing literature on the evolution of humans and other species that is too voluminous to review here. Suffice to point out however, that recent books by Spencer Wells and his associates, the Journey of Man, and Deep Ancestry, have been very influential in shaping views of recent human evolution for a general audience as well as practitioners of science. Two $DVDs$, Digital Versatile Disc or Digital Video Disc, based on the ideas contained in these books as well as other sources, have been shown on PBS, Public Broadcasting Service, television channels as well as National Geographic channels and have reached audiences in the millions.

A theme common to both these video presentations is that modern man descends from waves of individuals, consisting of small bands, who migrated out of Africa beginning about 60,000 years ago. Descendants from some of these bands eventually reached Australia and were the ancestors of aborigines presently living on that continent; while others reached the mountains of central Asia and remained for some time. Then, about 40,000 years ago, some bands from central Asia migrated eastward into what is now China;

while others migrated westward into Europe, where their descendants now live as well as on other continents such Australia, North and South America.

Given such theories as to the origins of modern humans, if some investigators were considering some genomic region as observed from a sample of individuals who were thought to be of European descent and who wish to check the plausibility of some of their ideas regarding the evolution of this genomic region using computer simulation, a good place to start a computer experiment would be about 40,000 years ago with an initial population consisting of a few hundred or so individuals. An example of the backward in time approach to simulating the evolution of a genomic sequence, using such ideas, is the paper by Schaffner *et al.* (2005). An interesting paper using the forward in time approach is that of Chadeau-Hyam *et al.* (2008) in which computer experiments based on the software package $FREGENE$ are reported. In both approaches, the idea of viewing the evolution of a genomic region from the perspective of a population bottle neck was used in both sets of computer simulation experiments. In this connection, it is of interest to point out the branching processes are well suited as a working paradigm for the computer simulation of populations evolving from bottle necks or small initial population as illustrated by the examples presented in chapters 10,11 and 12.

14.4 Statistics Used in Genome Wide Scans for Evolutionary Signals

The F_{ST} parameter will be defined in terms of a population that may be partitioned into $m \geq 1$ subpopulations. For example, if investigators were considering samples of human DNA from various geographic regions of the world such as some countries in Europe, Asia and Africa, then the population under consideration would be that of these three regions and the subpopulations would be those countries in Europe, Asia and Africa. Let A denote some genomic region under consideration, and let p_k denote the frequency of genomic region A in population $k = 1, 2, \ldots, m$. An example of a genomic region A could be a single SNP with p_k as the frequency of this SNP in subpopulation k. By way of illustration, note that $1 - p_k$ is the probability that a randomly selected individual in subpopulation $k = 1, 2, \ldots, m$ will not have the SNP in his DNA. The parameter F_{ST} will be defined in terms of the mean and variance of probabilities in the set $\{p_k \mid k = 1, 2, \ldots, m\}$.

By definition the mean of these probabilities is

$$\overline{p} = \frac{1}{m}\sum_{k=1}^{m} p_k \tag{14.4.1}$$

and their variance is

$$var\,[p] = \frac{1}{m}\sum_{k=1}^{m}(p_k - \overline{p})^2. \tag{14.4.2}$$

In this section, an operational definition of the parameter F_{ST} will be given as a function of the mean and variance defined in (14.4.1) and (14.4.2). Given these formulas, the parameter F_{ST} will defined as the ratio

$$F_{ST} = \frac{var\,[p]}{\overline{p}\,(1-\overline{p})}. \tag{14.4.3}$$

Apparently, this was the formula alluded by Nicholson *et al.* and other authors. If a reader is interested in a more detailed discussion of the parameter F_{ST} as well as other related parameters from the point of view of classical population genetics, it is recommended that the review paper by Excoffier (2003) be consulted.

The formalism just presented does not pertain directly to data. In what follows, data on a set of subpopulations will be an integral part in the estimation of the parameter in (14.4.3) as well as understanding the fundamentals underlying its nature. Suppose $m \geq 2$ subpopulations are under consideration and suppose that in subpopulation k, $k = 1, 2, \ldots, m$, the are $n_k \geq 1$ observations in the sample. By definition, let

$$n = \sum_{i=1}^{m} n_i \tag{14.4.4}$$

denote total sample size.

Before proceeding with the discussion, some further definitions will be required. If the sample consists of a readings of some property of the DNA from a sample of individuals, but the genotype of all individuals has not been ascertained, then the sample is said to be haploid, but, if the genotypes of all individuals have been ascertained, it will be said to be a sample of genotypic data. In the haploid case, the only information available is whether an individual carries some designated genomic region A, which may be a set of bases or a single base when a SNP is under consideration. In the case of a SNP, if an individual does not display the base at a particular location, he will be designated by the symbol a. From now on, attention will be focused on the case of haploid data, and to make the discussion more compatible with classical population genetics the genomic region A will be referred to as gene A or its allele a.

Let Y_{kj} be an indicator random variable for individual $j = 1, 2, \ldots, n_i$ in subpopulation $k = 1, 2, \ldots, m$. Then, by definition $Y_{kj} = 1$ if individual j carries gene A and $Y_{ij} = 0$ if this individual does not carry the gene A. Then, in subpopulation k, the number of carriers of gene A is the sum S_k

$$S_k = \sum_{j=1}^{n_k} Y_{kj}. \tag{14.4.5}$$

Observe that this sum is a random variable taking values in the set of non-negative integers

$$\{s \mid s = 0, 1, 2, \ldots, n_k\}. \tag{14.4.6}$$

The observed frequency of gene A in subpopulation i is, by definition,

$$\overline{p}_k = \frac{1}{n_i} S_k, \tag{14.4.7}$$

which is the arithmetic mean of the sample of indicators for subpopulation k.

The frequency $\overline{p}_k$ is a random variable so that a question that naturally arises is what is its expectation. At this point, the concept of probability introduced above plays an essential role. Let p_i denote the unknown probability that an individual selected at random from subpopulation i is of haplotype A. Then, as explained in chapter 1, the expectation of the indicator Y_{kj} is, by definition,

$$E\left[Y_{kj}\right] = 1p_k + 0p_k = p_k \tag{14.4.8}$$

for all $k = 1, 2, \ldots, m$ and for each k, $j = 1, 2, \ldots, n_k$. As the expected value operator is additive, the expectation of the observed frequency $\overline{p}_i$ is

$$E\left[\overline{p}_k\right] = \frac{1}{n_k} E\left[\sum_{j=1}^{n_i} Y_{kj}\right] = \frac{1}{n_k} \sum_{j=1}^{n_i} E\left[Y_{kj}\right] = p_k. \tag{14.4.9}$$

When attention is focused within the frequentist paradigm of statistics, the mean $\overline{p}_k$ is said to be an unbiased estimator of the unknown parameter p_k.

Let

$$\overline{p}_D = \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n_i} Y_{ij} \tag{14.4.10}$$

denote mean of the collection of indicators

$$D = \{Y_{kj} \mid k = 1, 2, \ldots, m; j = 1, 2, \ldots, n_k\}, \tag{14.4.11}$$

which are the observed data D. Another way of viewing this mean is that of the weighted mean of the means of the subpopulations as expressed in the formula

$$\overline{p}_D = \frac{1}{n}\sum_{k=1}^{m} n_k \overline{p}_k. \tag{14.4.12}$$

An analysis of variance in gene frequencies may based on the identity

$$Y_{kj} - \overline{p}_D = (\overline{p}_k - \overline{p}_D) + (Y_{kj} - \overline{p}_k). \tag{14.4.13}$$

By squaring and summing over all pairs k, j, it can be seen that

$$\sum_{k=1}^{m}\sum_{j=1}^{n_k}(Y_{kj} - \overline{p}_D)^2 = \sum_{k=1}^{m} n_k(\overline{p}_k - \overline{p}_D)^2 + \sum_{k=1}^{m}\sum_{j=1}^{n_k}(Y_{kj} - \overline{p}_k)^2. \tag{14.4.14}$$

Observe that this partition of the total sum of squares is valid, because

$$2\sum_{k=1}^{m}\sum_{j=1}^{n_k}(\overline{p}_k - \overline{p})(Y_{kj} - \overline{p}_k) = 0. \tag{14.4.15}$$

Due to some properties of indicators, the sum of squares on the right in (14.4.14) may be expressed in an interesting and useful form. At a first step, observe that

$$\sum_{j=1}^{n_k}(Y_{kj} - \overline{p}_k)^2 = \sum_{j=1}^{n_i} Y_{kj}^2 - \frac{1}{n_k}\left(\sum_{j=1}^{n_k} Y_{kj}\right)^2. \tag{14.4.16}$$

But, $Y_{kj}^2 = Y_{kj}$ for all pairs k, j. Hence,

$$\begin{aligned}\sum_{j=1}^{n_k}(Y_{kj} - \overline{p}_k)^2 &= \sum_{j=1}^{n_k} Y_{kj} - \frac{1}{n_k}\left(\sum_{j=1}^{n_k} Y_{kj}\right)^2 \\ &= n_k\overline{p}_k - n_k\overline{p}_k^2 \\ &= n_k\overline{p}_k(1 - \overline{p}_k)\end{aligned} \tag{14.4.17}$$

for all $k = 1, 2, \ldots, m$. Therefore,

$$\sum_{k=1}^{m}\sum_{j=1}^{n_k}(Y_{kj} - \overline{p}_k)^2 = \sum_{k=1}^{m} n_k\overline{p}_k(1 - \overline{p}_k). \tag{14.4.18}$$

By a similar argument, it can be shown that

$$\sum_{k=1}^{m}\sum_{j=1}^{n_k}(Y_{kj} - \overline{p}_D)^2 = n\overline{p}_D(1 - \overline{p}_D), \tag{14.4.19}$$

see, for example, the definition of $\overline{p}_D$ in (14.4.10). Finally, from the foregoing results, it can be seen that the partition of a sum of squares in (14.4.14) may be written in the form

$$n\overline{p}_D\left(1-\overline{p}_D\right)=\sum_{k=1}^{m}n_k\left(\overline{p}_k-\overline{p}_D\right)^2+\sum_{k=1}^{m}n_k\overline{p}_k\left(1-\overline{p}_k\right), \tag{14.4.20}$$

or equivalently

$$\overline{p}_D\left(1-\overline{p}_D\right)=\frac{1}{n}\sum_{k=1}^{m}n_k\left(\overline{p}_k-\overline{p}_D\right)^2+\frac{1}{n}\sum_{k=1}^{m}n_k\overline{p}_k\left(1-\overline{p}_k\right). \tag{14.4.21}$$

The first term on the right

$$var_{Among}\left[p\mid D\right]=\frac{1}{n}\sum_{k=1}^{m}n_k\left(\overline{p}_k-\overline{p}_D\right)^2 \tag{14.4.22}$$

may be designated as the weighted variance in gene frequencies among the subpopulations, given the data D. Similarly, given the data D, the second term on the right in (14.4.21)

$$var_W\left[p\mid D\right]=\frac{1}{n}\sum_{k=1}^{m}n_k\overline{p}_k\left(1-\overline{p}_k\right) \tag{14.4.23}$$

may be designated as the weighted sum of the variances in the gene frequencies within the subpopulations. Finally, the total variance in gene frequencies in the combined sample of subpopulations, given the data, may be denoted by

$$var_T\left[p\mid D\right]=\overline{p}\left(1-\overline{p}\right). \tag{14.4.24}$$

The identity in (14.4.21) is an analysis of variance in the sense that the partition,

$$var_T\left[p\mid D\right]=var_{Amog}\left[p\mid D\right]+var_W\left[p\mid D\right], \tag{14.4.25}$$

is valid. From the definition of the parameter F_{ST} in (14.4.3) the ratio

$$\widehat{F}_{ST}=\frac{var_{Among}\left[p\mid D\right]}{\overline{p}\left(1-\overline{p}\right)} \tag{14.4.26}$$

may be taken as an estimator of F_{ST}, given the data D. This is the estimator of F_{ST} mentioned by Balloux *et al.* (2002) as well as Lewontin and Krakauer (1973).

It is also of interest to define an estimator of F_W, given the data D, by the ratio

$$\widehat{F}_W=\frac{var_W\left[p\mid D\right]}{\overline{p}_D\left(1-\overline{p}_D\right)}. \tag{14.4.27}$$

Note that these ratio estimators have the property

$$1 = \widehat{F}_{ST} + \widehat{F}_{W} \tag{14.4.28}$$

for any sample D of indicators. Moreover, it also follows that

$$0 \leq \widehat{F}_{ST} \leq 1 \tag{14.4.29}$$

for all samples D. This result suggests that the estimator in (14.4.2) is more desirable than that used by Akey *et al.* (2002), which was based on some work of Weir and Cockerham, see Weir and Hill (2002), because their estimator of F_{ST} had the undesirable property that in some samples negative values could arise. Another advantage of the estimator in (14.4.25) is that it does not require the assumption that the gene frequencies of the m subpopulations are equal as did that used by Akey *et al.* (2002). With regard to the statistical performance of the estimator in (3.23), its sampling properties may be investigated by Monte Carlo simulation experiments as discussed by Nicholson *et al.* (2002) and others. From a purely statistical point of view, it is of interest to observe that whatever genetic interpretation is attached to the parameter F_{ST}, the estimator $\widehat{F}_{ST}$ is the fraction of the total variance in gene frequencies in the population as a whole to that is attributable to variance in gene frequencies among the m subpopulations.

Akey *et al.* (2002) have applied the F_{ST} statistic extensively in their interrogation of a high-density SNP map for signatures of natural selection in the human genome. In particular, they reported on an analysis of 26,530 single nucleotide polymorphisms ($SNPs$) with allele frequencies that were estimated in three populations. As a measure of genetic differentiation, F_{ST} was estimated for each locus and its distribution was estimated for the entire genome, each chromosome and individual genes. By doing backward in time simulations under the assumption of constant population size, the distribution of F_{ST} was also estimated under the hypothesis of neutral evolution and compared with the observed distributions. Many of the observed distributions of F_{ST} were not compatible with the hypothesis of neutral evolution, and 174 candidate genes, genomic regions, whose distribution of genetic variation, according to the authors, pointed to the action of natural selection in these genomic regions.

Other test statistics have also been applied in searching the human and other genomes for signatures of natural selection or for evidence for genomic regions implicated in the resistance or susceptibility to disease. In chapter 3, it was shown that if a large population evolves for an extended period with respect to one locus with many alleles, under the assumption of random mating and no mutation or selection, an equilibrium in gene and genomic frequencies is reached rather rapidly, which is known as a Hardy-Weinberg equilibrium. Such results have led researcher to develop statistical methods for testing the hypothesis that an observed sample is from a population in Hardy-Weinberg equilibrium-see chapter 3 for details. Such tests are often referred to under the name, Hardy-Weinberg disequilibrium, and if one types this name into a search engine for the internet, a large number of web sites will often appear. One such search yielded 196,000 sites that had this name in their title. There is an extensive statistical literature on the subject that is too large to review here, but if a reader is interested, it is suggested that he consult the internet for a preliminary search for references on this subject.

In chapter 3 it was also shown that, under the assumption of no mutation or selection and random mating, a large population would evolve to a state of equilibrium such that alleles at two or more loci were associated independently in the probabilistic sense even for loci located on the same chromosome. A population in such a state is said to be in linkage equilibrium, and if a population is not in this state, it is said to be in linkage disequilibrium. As the detection of linkage disequilibrium in a genome wide scans is thought to be an indicator of the action of natural selection or other evolutionary forces, there is an extensive statistical literature on this subject. For example, if one types the name, linkage disequilibrium, into a search engine for the internet, it can be seen that a large number of sites will have this name in their title. One search yielded 549,000 such sites devoted to lectures as well as scholarly articles on this subject. An interesting paper concerned with a comparison of linkage disequilibrium measures for fine-scale mapping of loci on a chromosome is that of Devlin and Risch (1995). A more recent review article on linkage disequilibrium in humans and the models and data used to detect it is that of Pritchard and Przeworski (2001). An example of a paper in which tests for linkage disequilibrium were applied in genome wide scans is that of Hinds *et al.* (2005). An extensive list of papers in which tests for linkage disequilibrium were applied may be found in the on line supporting material of the paper by Sabeti *et al.* (2006). From the point of view of medical applications of such results, an interesting note on the harvesting of medical information from the human genome is that of Altshuler and Clark (2005).

From the point of view of statistics, the wide spread application of such high through put devices as microarrays, in which many statistical hypotheses may be tested simultaneously, has made mandatory the development of procedures to control error rates arising from the consideration of many tests. For example, if one supposed that all hypotheses were false, then by pure chance in a large number of tests some of the hypotheses may be declared statistically significant. Among the papers that have been influential in proposing methods for controlling error rates, when many hypotheses are under consideration, is that of Benjamini and Hochberg (1995), which is now known as the False Discovery Rate (FDR) method. More recent papers on extending this method are those of Benjamini and Yekutieli (2001) and (2005). In a more recent paper, Efron (2008) has developed these methods further by considering empirical Bayes and the two-groups model in the context of experiments using microarrays. Also included in this paper are four specific examples in which the methods proposed in this paper would be applicable. Included in the topics discussed in this paper are the choice and meaning of null hypotheses in large-scale testing situations, power calculations, the limitations of permutation methods, significance testing for groups of cases, correlation effects, multiple confidence intervals and Bayesian competitors to the two-groups model.

Hoggart *et al.* (2008) have also considered genome-wide significance testing for dense SNP and resequencing data. Among the methods mentioned by these authors is the FDR method and others, but the more innovative approach presented by these authors to finding solutions to problems that arise in genome-wide significance test is that of using the $FREGENE$ software package to simulate the evolution of 5 Mb regions of DNA in a diploid population, under various versions of the "Out of Africa Hypothesis" mentioned in a preceding section of this chapter and also in chapter 8. Such methods appear to be very promising, but before they can become acceptable to the community of applied mathematics as a whole and particularly to those who are familiar with the concepts used in stochastic processes and statistics, the algorithms underlying such simulations will have to thoroughly documented so that their merits and demerits may be examined by interested researchers in these fields.

14.5 Bayes Factors and Other Measures Used in Detecting Signals of Natural Selection

Interestingly, Grossman *et al.* (2010), reported on a statistical procedure called CMS, composite multiple signals, for distinguishing causal variants in genomic regions of positive selection. As work on the human and other genomes have progressed over the last decade, hundreds of selectively positive genomic regions have been identified. Typically, these regions are large and may contain anywhere between hundreds of kilobases up to megabases with many genes and thousands of polymorphisms. Using data from the International Haplotype Map project, these authors reported that they were able to localize population specific signals down to a median of 55 kilobases, which included known and novel forms, by using the CMS procedure. Unfortunately, from the point of view of workers in stochastic processes and statistics as well as other readers who are interested in statistical methodology, the presentation of the formal basis for CMS in the supporting online material is very sparse; consequently, the purpose of this section is to present a formal account of some background concepts that seem to underlie this statistical procedure. It should be stated at the outset, that the ideas which follow are interpretations by the senior author of the ideas presented by the authors in the papers under consideration. These ideas may or may not coincide with those of the authors of these papers. In the remainder of this section, some background material will be discussed in preparation for a discussion of the Grossman *et al.* (2010) paper, which will be presented in more depth in the next section.

One of the concepts mentioned in the supporting material in Grossman *et al.* was a ratio called the Bayes Factor. When this term is typed into a search engine on the internet, it can be seen that there is an extensive literature on this subject. A lengthy tutorial on Bayes Factors has been given by Kass and Raferty (1993). Suppose an investigator has an observed set a data D and it has been postulated this data has been generated according one of two alternative mechanisms stated in terms of two hypothesis H_1 and H_2. An investigator may have prior opinions as to how the observed data were generated, which may have been stated in the formalized form of a mathematical model or some verbal rhetoric concerning these two hypotheses. At this point, it will be assumed that an investigator quantifies his prior opinions regarding the two alternative hypotheses in terms of personal prior probabilities $P[H_1]$ and $P[H_2]$ such that $0 < P[H_1] < 1$ and $P[H_2] = 1 - P[H_1]$. Let $P[D \mid H_1]$ denote the conditional probability of the data D, given that hypothesis H_1 was in force and define the conditional probability $P[D \mid H_2]$ similarly. Then, by an application of Bayes's theorem, the posterior probabilities of hypothesis H_k, given the data D, is

$$P[H_k \mid D] = \frac{P[H_k] P[D \mid H_k]}{P[D]}, \tag{14.5.1}$$

for $k = 1, 2$ where

$$P[D] = P[H_1] P[D \mid H_1] + P[H_2] P[D \mid H_2]. \tag{14.5.2}$$

In Bayesian statistics, the relative plausibility of the two hypotheses is judged in terms of the odds ratio, which will be denoted by *PostOdds*. For the case under consideration this ratio is defined as

$$PostOdds = \frac{P[H_1 \mid D]}{P[H_2 \mid D]} = \frac{P[H_1] P[D \mid H_1]}{P[H_2] P[D \mid H_2]}. \tag{14.5.3}$$

Observe that $P[H_1 \mid D] + P[H_2 \mid D] = 1$. The ratio

$$BF = \frac{P[D \mid H_1]}{P[D \mid H_2]} \tag{14.5.4}$$

is called the Bayes Factor. The posterior odds ratio (14.5.3) has an interesting verbal interpretation; namely the

$$\text{posterior odds} \ = \ \text{prior odds} \times \text{Bayes Factor}.$$

It is also interesting to note that if the prior probabilities of the two hypotheses are equal, $P[H_1] = P[H_2] = 0.5$, indicating that an investigator is neutral as to which hypothesis is acceptable, then the posterior odds ratio coincides with the Bayes Factor. Thus,

$$PostOdds = \frac{P[H_1 \mid D]}{P[H_2 \mid D]} = \frac{P[D \mid H_1]}{P[D \mid H_2]} = BF. \tag{14.5.5}$$

Evidence for and against the hypothesis H_1 is judged in terms of values of the posterior odds ratio, which in the neutral case is BF. As can be seen from (14.4.5), the range or set of possible values of this ratio in the interval $(0, \infty)$. If $P[D \mid H_2]$ is small relative to $P[D \mid H_1]$, then BF would be large, and this large value would be interpreted as evidence in favor of H_1. On the other hand, small values of BF would be interpreted as evidence in favor of H_2. If a reader is interested in further details, regarding the interpretation of an observed value of BF, Kass and Raftery (1993) may be consulted for further details. Given a sample of data from the International Haplotype Map project and a set of $SNPs$ to be considered, a first step is to compute some standardized scores, which are functions of the observed data as well as simulated data that will be used to test hypotheses for selection. Voight *et al.* (2006) have provided an informative discussion of the concepts and methods used in computing these scores. Their procedure begins with the extended haplotype homozygosity (EHH) statistic proposed by Sabeti *et al.* (2002). The EHH statistic is a measure of the decay of identity, base similarity, as a function of the distance, expressed in terms of some unit such as a Mb, from a core SNP, allele, under consideration. This function starts at 1 near the SNP and decreases as a function of distance from the origin. If a genomic region marked by a SNP has been under strong positive selection for a long period of time, then one would expect to find extensive haploid homozygosity in the sense that individuals shared bases so that the decay of base similarity would be slower than if no selection were acting on the region. In principle, realizations of this decay function may be simulated under the hypothesis of neutrality, no selection, or under the alternative hypothesis that positive selection has been acting on the region of DNA, starting from some time point in an assumed evolutionary history of the population from which the sample of individuals under consideration was collected. Evidently, one may also include the effect of mutations on the simulated function.

There are also other notions used by Voight *et al.* (2006) in their calculations when considering a SNP and its allele. These notions are called ancestral and derived, see Figure $1A$ of their paper for a graphical representation of these notions. In Figure $1B$ there are simulated graphs of EHH decay functions for ancestral and derived alleles, which were computed under the alternative hypotheses of no selection and selection. As can be seen from this figure, these graphs are decreasing functions of the distance from the origin, which is the location of the SNP. The notions of ancestral and derived alleles were not explicitly defined by the authors, but as a working definition of these ideas, an allele with higher frequency in a population will be designated as ancestral; while the allele with a lower frequency will be called derived. Presumably, derived alleles have arisen by the mutational process of nucleotide substitution and are "younger" than the ancestral allele.

The data set used by Voight *et al.* (2006) consisted of about 800,000 polymorphic $SNPs$ in a total of 209 unrelated individuals. In particular, this sample consisted of 89 individuals from Beijing and Tokyo which were referred to as east Asian, 60 individuals of western European origin and 60 individuals Yoruba from Ibadan, Nigeria. In terms of the notation used in section 14.5.5, there are $m = 3$ subpopulations, with sizes $n_1 = 89, n_2 = 60$ and $n_3 = 60$ for a total of $n = 209$ individuals. It should be mentioned in passing that if the DNA of each individual were sequenced and aligned among the individuals in a sample, then the EHH decay functions for the ancestral and derived alleles could, in principle, be estimated for each SNP under consideration in each subpopulation by observing the decline in the number of matches in the bases among individuals as a function of the distance from the origin.

If one computes the area under two EHH curves, which have been computed under the alternative hypotheses of selection or no selection, then, because under the hypothesis that selection has occurred the curve declines much more slowly than in the neutral case, one would expect to find that the area under curve would be greater than that for the case of no selection. For a genomic region under consideration, the SNP may occur at the start of a region or the SNP may be in a central point of the region. To compute the area under a curve, one computes an integral, which will usually be approximated by a summation process. In all cases reported by the authors, the integral would be computed from the point from where the EHH curve was one to the point where it reached 0.05. For each case, integrals were computed for both the ancestral and derived alleles. Let the symbols iHH_A and iHH_D, respectively, denote the integrals or area under the EHH curve for the cases in which is was assumed that an allele was ancestral or derived. Then, the first step in the computation of a score for the genomic region k under consideration was to compute the logarithm

$$R_{jk} = \ln\left(\frac{iHH_A}{iHH_D}\right) \tag{14.5.6}$$

for genome region k in subpopulation j. Observe that, because R_{jk} is the log of a ratio, it is a pure number in the sense the unit of area calculated in the integrals cancels out in the ratio iHH_A/iHH_D. In theory, the set of possible values of R_{jk} is the set of real numbers $(-\infty, \infty)$ and if $iHH_A/iHH_D \approx 1$, then $R_k \approx 0$. Large negative values of R_k are indicative of long haplotypes carrying the derived allele; whereas large positive values of R_k are indicative of long haplotypes carrying the ancestral allele.

The next step in the computation of a score for a genomic region k in subpopulation j is to standardize it by using the following computations. In the general case of $m \geq 2$ subpopulations, define a probability distribution by letting

$$p_j = \frac{n_j}{n} > 0 \tag{14.5.7}$$

where

$$n = \sum_{j=1}^{m} n_j \tag{14.5.8}$$

Then, by definition, the expected score for genomic region k corresponding to some SNP for the population as a whole is

$$E\left[R_k\right] = \sum_{j=1}^{m} p_j R_{jk}. \tag{14.5.9}$$

Similarly, the variance in genomic scores for genomic region k among the subpopulations is, by definition,

$$var[R_k] = \sum_{j=1}^{m} p_j \left(R_{jk} - E\left[R_k\right] \right)^2 \qquad (14.5.10)$$

and the standard deviation for this region is $sd\left[R_k\right] = \sqrt{var[R_k]}$. Given these definitions, the standardized score for genomic region k in subpopulation j is defined as

$$stadR_{jk} = \frac{R_{jk} - E\left[R_k\right]}{sd\left[R_k\right]}. \qquad (14.5.11)$$

Observe that $stadR_{jk}$ has expectation 0 and variance 1. It should be stated that the distribution used in the calculation of $stadR_{jk}$ may not be the same as that referred to in formula (2) of Voight *et al.* (2006). However, because the verbal description of the distribution used by these authors was too vague to express in mathematical terms, the distribution used above may be viewed as a working definition. In practice, an investigator is free to choose any distribution that is judged to fit his purposes, but, at the some time, the construction of this distribution should be described in precise mathematical terms. In their data analysis, the authors in Voight *et al.* (2006) calculated $stadR_{jk}$ for every SNP with a minor allele frequency ≥ 0.05. Each value of $stadR_{jk}$ at each SNP was viewed as a measure of the strength of the evidence of natural selection acting at the SNP or in a genomic region near the SNP. If a reader is interested in further details, it is suggested the paper be read in greater depth.

14.6 Applications of Simulated Genomic Data in the Evaluation of Statistical Tests

One of the interesting aspects of the methodology used by Grossman *et al.* (2010) was that of applying a model to simulate the evolution of genomic regions containing up to one million bases, $1Mb$. The output of this simulation model was then used to evaluate the performance of the various statistical procedures considered by the authors to detect signals of positive selection. The simulation model used by the authors was described in some detail in Schaffner *et al.* (2005). Basically, it is a coalescent model, which looks backward in time, and simulates human genome sequence variation. The authors state that simulation software was validated by comparing its predictions with those of a standard Wright-Fisher models, which were treated in detail in chapters 4 and 5. It is stated that software accommodates an extensive range of demographic histories for multiple populations, which include changes in population size, bottle necks and migrations as well as a simple gene conversion model and variable rates of genetic recombination. The simulation software package may be obtained from the website $http : //www.broad.mit.edu/\tilde{\ } sfs/cosi$ and is sometimes referred to as $\cos i$. In principle, by reading the code in this software package, it would be possible to deduce the structure of the mathematics underlying the code, but this could be a very tedious process. As an aid to making the basis of the calculations transparent to the mathematical community, it would have been very helpful if the authors had provided a mathematical description of algorithms underlying the software by collaborating with someone in that community. Had a formal account of such an exercise been presented, the interesting and potentially important ideas described in what follows could be studied with a higher level of credibility by those in the mathematical community.

To test each statistical procedure's ability to localize recent signals of positive selection and to distinguish variants thought to be causal from nearby neutral markers, neutrally evolving genomic regions and region regions containing a positively selected allele were simulated using $\cos i$. A range of so-called demographic models were tested in these simulation experiments. Included in these models were a standard model, calibrated models for European, East Asian and West African populations as well as several more extreme models. Genomic regions under selection were modeled in terms of a single centrally located variant that was supposed to have arisen by mutation within the last 5,000 to 30,000 years and was subject to a specified intensity of selection such that it rose to present day frequencies ranging from 0.2 to 1. Evidently, selection was characterized in terms of selection coefficients for the two allele case discussed in chapter 4. In each computer simulation experiment, characterized by assigning numerical values to parameters for a population under consideration,1500 replications were simulated and each replicate consisted of $1Mb$ of simulated sequence data with about 10,000 polymorphisms for 120 chromosomes from each population. In addition a simulated data set was generated that matched the frequency distribution and density of Phase II of the International Haplotype Map Project, see K. A. Frazer *et al.* (2007) International HapMap Consortium, Nature 449: 851 and the internet for more details.

Although the authors did not used the word "stochastic" the computation of this many replications suggests that they were computing realizations of a stochastic coalescent processes that was, in some sense, related to the Wright-Fisher processes used in the experiments presented in chapter 5, which were forward in time Monte Carlo simulations. The computation of these many replications would entail many calls to a random number generator which could exceed its period, which raises a question as to the validity of the simulated random numbers used in their experiments. Chapter 5 may be consulted for a discussion of the period and other properties of random number generators. Even though the random generator that was evidently used in their computer simulations may be questioned, nevertheless, the idea of using simulated genomic data to test the sampling properties of proposed statistical tests is important and deserves further intensive development in the coming years as sequenced DNA data becomes more plentiful in the public domain.

Altogether five statistics were used that had distinguishable distributions for causal and neutral variants, which included neutral variants in regions in or near regions of positive selection. Among these statistic was F_{ST} that was discussed in a previous section as well as those discussed in section 14.5.5. However, in what follows, attention will be devoted CMS, a statistical framework that involves, among other things, ideas from Bayesian statistics. Within this framework, many measures for evidence of positive selection may be considered. For the sake of brevity, only two statistical procedures will be considered here. One of these procedures was labeled the iHS test by the authors, and the other was labeled the ΔiHH test.

The iHS test follows along lines similar to that used by Voight *et al.* (2006) as discussed in section 14.5. In fact, the iHS statistic coincides with the statistic defined in equation (14.5.11) and is standardized in the sense described in the derivation of this equation. The authors state that this test was performed only for bi-allelic $SNPs$ whose minor allele frequency was above 0.05. As discussed in section 14.5, in these calculations, A denoted the ancestral allele and D the derived allele. Evidently, for the simulated data, the notions of ancestral and derived could be defined in terms of the model, but, for real data, these terms were based on evolutionary hypotheses that will be mentioned below. For all chromosomes carrying the allele A, the EHH scores were calculated between the core SNP and every bi-allelic SNP to within $2Mb$. The area under the EHH curve was integrated with respect to the genetic distance expressed in terms of centimorgens, cM, by linearly interpolating between successive bi-allelic $SNPs$, until the EHH curve dropped to 0.05. If this curve did not drop to 0.05 within $2Mb$, then the iHS was dropped for this SNP. The integral for the ancestral and derived alleles were then calculated and denoted respectively by iHH_A and iHH_D. As explained in section 14.5, the unstandardized scores were then calculated as $\ln(iHH_A/iHH_D)$.

In the analyses of simulated data, unstandardized scores were calculated for every bi-allelic SNP in each simulated population. These unstandardized scores were then partitioned into 20 equally sized bins according to their so-called derived allele frequencies. Within each bin, for the case of neutral simulation, all scores were standardized so that they had mean zero and variance one, and then, evidently, the same distribution as that used to standardize scores in the neutral case was again used in an effort to "normalize" the scores in all other simulation scenarios. It should be mentioned that the above description of the normalization process for each bin was an attempt to deal with the jargon, consisting of many undefined terms, used by the authors in terms of standard concepts widely used in statistics, but this description may or may not correspond to the actual data manipulations performed by the authors.

For the analysis of the $HapMapII$ data set, information as to the ancestral state of each allele was based information in the chimpanzee and macaque genomes. If it was available, the ancestral allele was taken as the corresponding base in the chimpanzee genome, if it were not present in the chimpanzee genome, the corresponding base in the macaque genome was used. If neither base were available, no ancestral state was inferred.

For the test labeled ΔiHH the integrals iHH_A and iHH_D were again computed for every core SNP being considered. The unstandardized ΔiHH scores were then computed as the absolute difference $\mid iHH_A - iHH_D \mid$. These absolute differences were then calculated for each bi-allelic SNP in a putative selected population. In the analysis of the simulated regions as well as the $HapMapPhaseII$ data, the unstandardized scores were sorted into 20 equally sized frequency bins and then standardized, by some uniform procedure, to have mean zero and variance one. It will be noted that ΔiHH captures the magnitude of haplotype length whereas iHS captures the relative sizes of the ancestral and derived haplotypes. For a more in depth discussion of the differences in these two types of scores, the online supporting material provided by the authors may be consulted.

For each of the five tests considered by the authors an empirical distributions of the scores were estimated from simulated data on the evolution of a genomic regions containing $1Mb$ such that subregions of this region were either under selection or were evolving according to a neutral theory. For a more detailed discussion of the ideas of selective and neutral evolution within the Wright-Fisher paradigm, a reader may wish to consult chapters 4 and 5, where forward in time evolution was considered. It may also be on interest to consult chapters 10, 11 and 12, where selection and neutrality are defined within a paradigm of evolution characterized by parameterized branching processes. The authors estimated the empirical distributions from 1,000 neutral regions and 7,500 regions under selections as simulated by the stochastic coalescent model used in their studies, which was calibrated, or adjusted to, three *HapMapII* populations of West African, European and East Asian origin.

As all scores were standardized, the theoretical set of values for these scores would be the set $(-\infty, \infty)$ of real numbers. To help fix ideas, let the hypothesis H_1 denote the idea that selection was operative in the evolution for some genomic region and let H_2 an alternative hypothesis that a region had undergone neutral evolution. Let $P(s \mid H_1)$ denote the probability density function of the distribution of scores for some test under the hypothesis H_1 that selection was in force. Then the problem is to find a statistical estimation procedure that would estimate this density for every $s \in (-\infty, \infty)$. There is an extensive statistical literature on solutions to this problem, but rather than consulting this literature for guidance as to how to proceed with an estimation procedure, it seems prudent to consider the special nature of the problem at hand. If, for example, the scores were approximately normally distributed with expectation 0 and variance 1, then with high probability the sample of realized scores will lie in the interval $[-4, 4]$, but, even if the realized scores are do not follow a standard normal distribution, then is seems likely that they will all lie within some finite interval.

By way of an illustrative example, suppose a sample of realized scores belong to the interval $[-6, 6]$. Then, the problem of estimating the density $P(s \mid H_1)$ reduces to that of estimating a value of the density at every point $s \in [-6, 6]$. But, this reduced version of the estimation problem however, may still be too complicated so that it would seem preferable to find a simpler and more practical solution to the problem. Let S denote the sample of realized scores and suppose, for example, that the interval $[-6, 6]$ is partitioned into a set of disjoint intervals of the form $[-6, -5), [-5, -4), \ldots, [5, 6]$, where $[-6, -5) = (s \mid -6 \leq s < -5)$ and so on up to $[5, 6] = (s \mid 5 \leq s \leq 6)$. Furthermore, suppose there are $m \geq 2$ disjoint sets in this partition. In general, let $[s_\nu, s'_\nu)$ denote the interval $\nu = 1, 2, \ldots, m$ in this partition, and, under the hypothesis H_1 that section was in force in the simulated data, let $n[s_\nu, s'_\nu)$ denote the number of observed scores belonging to the interval $[s_\nu, s'_\nu)$. Then, for the test represented by the score under consideration, an estimate of the density $P(s \mid H_1)$ at the point s_ν would be assigned the value

$$\widehat{P}(s_\nu \mid H_1) = \frac{n[s_\nu, s'_\nu)}{7,500} \tag{14.6.1}$$

for $\nu = 1, 2, \ldots, m$ with the proviso that for the last set in this partition one would use the closed interval $[s_m, s'_m]$. On the other hand, if the hypothesis H_2 of neutral evolution were in force in the simulated data, then the estimate

$$\widehat{P}(s_\nu \mid H_2) = \frac{n[s_\nu, s'_\nu)}{1,000} \tag{14.6.2}$$

would be used in place of that in (14.6.1) for $\nu = 1, 2, \ldots, m$. Although the authors did not present any of the details regarding the procedure they used to estimate the probability densities under consideration, the estimators described in (14.6.1) and (14.6.2) are seen to be useful, because by choosing a suitable value of m, the number of sets in a partition, one could assure that the numbers $n[s_\nu, s'_\nu)$ would be sufficiently large to ensure that the above estimates of the densities under the two alternative hypotheses would be reliable.

Given the above non-parametric estimates of the probability density function, the next step in the statistical analysis procedure used by the authors was to use these estimates in the analysis of the actual data. Evidently, the real data used by the authors was sequenced. In general suppose there are $k \geq 2$ tests under consideration and recall in the analyses of the real data the authors included five tests, $k = 5$, and for illustrative purposes only two of these tests or types of measurement were discussed above. Let $(s_1, s_2, \ldots, s_k)$ denote the scores assinged to a genomic region near a *SNP* under study in the real data and suppose the k scores are distributed independently. Then, under the hypothesis H_1 that this genomic region had evolved under selection, the probability assigned to these scores is

$$\prod_{\nu=1}^{k} \widehat{P}(s_\nu \mid H_1). \tag{14.6.3}$$

Similarly, under the hypothesis H_2 that this genomic region had evolved according the postulates of neutral evolution, the probability assigned to the k scores would be

$$\prod_{\nu=1}^{k} \widehat{P}\left(s_\nu \mid H_2\right). \tag{14.6.4}$$

Given these assigned probabilities of observing the scores $(s_1, s_2, \ldots, s_k)$, a Bayes Factor

$$BF = \prod_{\nu=1}^{k} \frac{\widehat{P}\left(s_\nu \mid H_1\right)}{\widehat{P}\left(s_\nu \mid H_2\right)}. \tag{14.6.5}$$

could be calculated for each SNP under consideration. The CMS score for each SNP is then, by definition, the posterior conditional probability of H_1, given the data $(s_1, s_2, \ldots, s_k)$, which was calculated using the following formula. Let $P\left[H_1\right]$ denote the prior subjective probability or prior weight given the hypothesis H_1 of selection and let $P\left[H_2\right] = 1-\ P\left[H_1\right]$ be the subjective prior probability for H_2 of neutral evolution. Then, given the score s_ν for test ν, the posterior conditional probability of H_1 is given by the formula

$$P\left[H_1 \mid s_\nu\right] = \frac{P\left[H_1\right] \widehat{P}\left(s_\nu \mid H_1\right)}{P\left[H_1\right] \widehat{P}\left(s_\nu \mid H_1\right) + P\left[H_2\right] \widehat{P}\left(s_\nu \mid H_2\right)} \tag{14.6.6}$$

for $\nu = 1, 2, \ldots, k$. Then, under the assumption that the k tests are independent, the score CMS for a particular SNP is the posterior probability

$$CMS = P\left[H_1 \mid (s_1, s_2, \ldots, s_k)\right] = \prod_{\nu=1}^{k} P\left[H_1 \mid s_\nu\right]. \tag{14.6.7}$$

Let N_{SNP} denote the number of $SNPs$ under consideration. Usually, this number is rather large. The authors have suggested that the value $P\left[H_1\right] = 1/N_{SNP}$ be used for all $SNPs$ being considered. Then, $P\left[H_2\right] = 1 - 1/N_{SNP}$ so that under this assignment of prior probabilities, the hypothesis of neutral evolution H_2 would be favored. Presumably, with this approach the null hypothesis of neutral evolution would be in force of all $SNPs$ being considered so that the outliers in the empirical distributions of BF and CMS computed from all SNP would be indicators that the process of selection was operative during the evolution of a genomic region associated with the outlier $SNPs$. Presumably, other prior evolutionary hypotheses could be incorporated into the prior probabilities, but these ideas will not be pursued here. It suggested, however, that an interested reader consult the supportive online material given by the authors where these and other issues such as False Discovery Rates are discussed.

14.7 An Overview of Data and Software for Constructing Species and Gene Trees From Mammalian Genomic Data

In the preceding sections of this chapter, attention has been devoted to statistical methods for detecting regions in the human genome where it is thought that natural selection has acted in the recent past. In terms of human evolution, recent past refers to periods of time ranging from 10,000 to 100,000 years, which covers the period of time that ancestors of modern humans migrated out Africa and the subsequent development of agriculture in various regions of the world. When the genomes of several mammalian species are under consideration however, and regions in their genomes are being tested for evidence of natural selection, the periods of evolutionary time involved in such investigations are often expressed in millions of years, representing the time two or more species are thought to have diverged from a common ancestor.

On the other hand, when regions in the genomes of domestic species of plants or animals are being tested for evidence of artificial selection, which is thought to have occurred in connection with the development agriculture and the formation of urban areas, then shorter periods of evolutionary time consisting of 10,000 or so years are under consideration. In recent years, the genomes of a number of mammalian species have been sequenced so that it is now possible to search the genomes of a number of species for evidence of natural and artificial selection. There is a large literature molecular evolution as alluded to in chapter 6, but in this section, attention will be focused on an overview of a recent $Ph.D.$ thesis, Raj (2009), devoted to defining and testing for signatures of natural and artificial selection in mammalian genomes.

Homology is one of the most important concepts in evolutionary biology and, at the phenotypic level, refers to the similarity of characteristics among individuals of one or more species, which are thought to have a common ancestor. At the molecular level, homology among proteins and DNA is often assessed by either high similarity of DNA sequences, or, for the case of proteins, similarity in their sequences of amino acids. Genes or proteins that are related by homology are called homologs, which are in turn subdivided into two types, orthologs and paralogs.

If, for example, a species diverges into two distinct species, the divergent copies of a single gene in the resulting two species are said to be orthologs. Paralogs, however, are homologs that are separated by a gene duplication event. For example, if a gene is duplicated such that it occupies two regions of the same genome, then the two copies are called paralogs, which in turn are subdivided into two classes. The term, in-paralog, is applied to those cases in which duplication occurred after two species had evolved from a common ancestor as a result of a speciation process; whereas out-paralogs refers to those cases in which gene duplication occurred before speciation. Gene duplication is a type of mutation process and models of this process should be included in stochastic computer simulation models designed to simulate the evolution of molecular DNA.

The data used by Raj (2009) consisted of amino acid and DNA sequences that were downloaded from data bases in the public domain. The amino acids sequences and their corresponding coding DNA sequences (CDS), represented by the largest transcript of each ortholog, were retrieved from the Ensembl core data bases. Only "known" Ensembl protein coding genes were included in the data set, which had been put through rigorous confirmatory procedures which will not be discussed in this overview. Pseudogenes, $mtDNA$ genes and $miRNA$ genes were all excluded from the ortholog data set. At this point in the discussion, a reader may wish to consult the material on the definition of gene presented in chapter 13.

Data on sequenced genomes of fourteen mammalian genomes were used in the thesis and nine of these genomes were said to be of high coverage. The term, high coverage, refers to the extent the reported DNA sequences actually match the true DNA of the individual or individuals from whom the sample of DNA was drawn. For example, if the phrase "8-fold coverage of sequenced genomes" is entered into a search engine on the internet, then thousands of sites pertaining to this phrase may appear. Included in these sites are descriptions of the algorithms, based on graph theory, that are used to assemble short strips of DNA into long sequences, but a description of the details involved in these algorithms is beyond the scope of this book.

Suffice it to state that out of the 14 genomes considered in the thesis, nine were high-coverage genomes. Specifically, these nine genomes included the human, chimpanzee, orangutan, rhesus macaque, mouse, rat, dog, cow and horse. Among mammals, the human and mouse genomes have been sequenced most extensively with each base being represented by at least a 8-fold coverage. A small group of mammals, cattle, dogs and horses, which are important to agricultural, medical and evolutionary research, have also been targeted for extensive high coverage genome sequencing.

The orthologous relationship among the species for each gene was inferred from two prediction algorithms named the Ensembl Compara and the Online Codon-Preserved Alignment Tool (*OCPAT*), see Raj (2009) for references. The first set of orthologs was obtained from the Ensembl Compara multiple species data base, which uses an automatic pipeline of syntenic whole genome alignments and the best reciprocal *BLAST* (Basic Local Alignment Search Tool) hit matches. The Ensembl Compara data base is based on a comprehensive computational pipeline to handle orthology and paralogy gene prediction, clustering, multiple alignment and tree generation, which accommodate the handling of large gene families.

The gene orthology and paralogy prediction pipeline has seven basic steps, but, for the sake of brevity and Illustative purposes, only two of these steps will be briefly described. For example, in step 6 from the aligned cluster or protein sequences build a phylogenetic tree using a software package called TreeBeST, which uses a back-translation based on the amino acids of the protein to the coding *DNA* sequence under consideration. Then, step 7 consists of inferring gene pair-wise relations of orthology and paralogy from the phylogenetic tree.

At this point in the discussion, a reader may have the impression that a great deal of processing of the *DNA* sequence data as well as the sequenced protein data must be carried out before the task of detecting signals positive selection in the genomes of the species under consideration can begin. In what follows, only a very brief outline of the ideas used processing the data. One of the tools, which was called the Online Codon-Preserved Alignment Tool (*OCPAT*). *OCPAT* extracts putative orthologous genes from multiple genomes and aligns the orthologs with the reading frame maintained in all species. *OCPAT* then determines *CDS* regions by alignment of the longest representative $\operatorname{Re} f - Seq\ mRNA$ for each gene to the human genome. *CDS* alignments were then extracted, which represented the 14 eutherian mammals under study. The *OCPAT* pipeline has five steps, which will not be discussed in this overview.

In any comparative genomic study statistical methods are applied to thousands of genes many of which will be difficult to align, because of highly diverse sequences or sequencing errors. Therefore, in most statistical analyses of genomic data the uncertainties in multiple sequence alignments are not taken into account. In an attempt to address these issues of uncertainty two controls were used to test the accuracy of two alignment algorithms. Each of these procedures were used on the phylogenetic branch for the domestic horse, which was the most divergent species in the study. It turned out that output of the alignment software packages $T - COFFEE$ and $MUSCLE$ were highly correlated in that the measure labeled $\Delta \ln l$ for the two outputs had the correlation coefficient $r = 0.91$ with a p value of $P < 10^{-21}$, indicating that the observed correlation was statistically significantly different from 0. This test did not completely address uncertainty issues that arise in the alignment of DNA sequences but it does suggest that the results are robust.

The next issue that needs to be addressed in detecting signals of natural selection that arise from long evolutionary time periods is that of codon alignment among the species under consideration. The most optimal and reliable approach to constructing codon alignments is to use a protein alignment and then back-translate this alignment into a codon-based DNA alignment. A modified version of the Phylogenetic Analysis by Maximum Likelihood program ($PAML$) called $PAL2NAL$ was used to convert multiple sequences of protein alignments and the corresponding DNA sequences into a codon alignment.

It should be noted that this program takes alignments even if the input DNA sequence has mismatches with the input protein sequence or even if it contains an untranslated region and poly A tails. The use of this program entails three steps that will not be discussed here.

After a set of tentative orthologs were extracted from the data, the sequence alignments were subjected to stringent quality control. To implement these quality control measures stringent filters were designed to avoid mismatches and frameshifts, minimize the impact of annotation errors and enhance sequence quality. After an initial screening of the orthologs, the codon-based alignments were further refined with a program called *GBlocks*. As an aid for the reader to come to grips with the quality control measures used in the research, the *GBlocks* analysis was carried out with the following stringent options: (1) the minimum length of an alignment block was set to 10 positions; (2) all gap positions were excluded; and (3) the maximum number of contiguous non-conserved positions allowed was set to 4.

The next step in the analysis of the data aimed at detecting signals of long term natural selection was that of constructing phylogenetic trees. The Maximum Likelihood (ML) tree for each ortholog cluster was reconstructed using a program called $TreeBest$ release 2.4.5. Evidently, a related program called $TreeBeST$ merges several trees that were constructed from the same alignment data using different methods such as a Neighbor-Joining (NJ) and maximum likelihood. $TreeBeST$ takes as an input a species tree and attempts to build a gene tree that is consistent with the topology of the species tree. This species-guided approach allows the gene tree to have a topology that is consistent with the species tree, with the proviso that the alignment data supports this view.

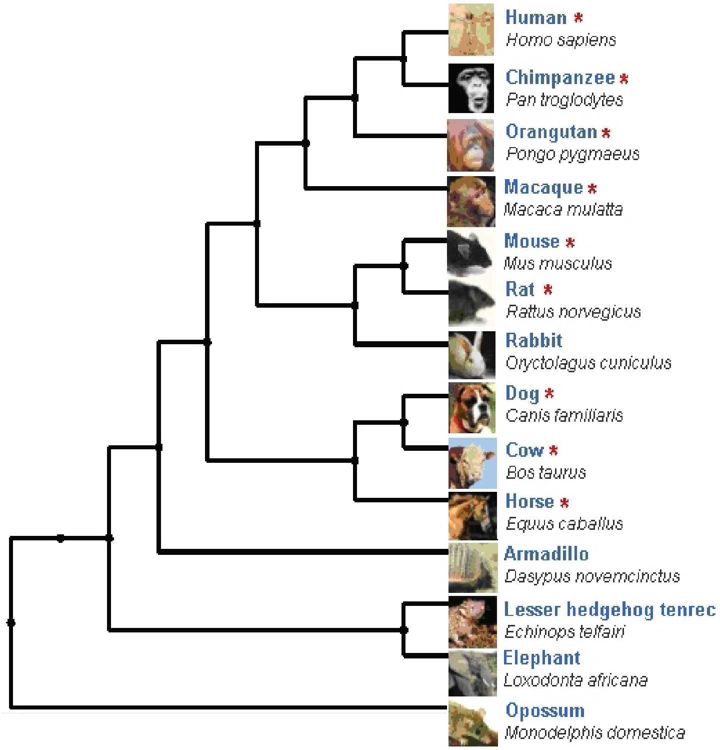

Figure 14.7.1 The Consensus Tree for Phylogenetic Reconstruction.

Altogether 5 trees were built, using various methods that will not be described here, and a final tree or consensus tree was built from these 5 trees using the "tree merging algorithm". This procedure allows $TreeBeST$ to take advantage of the fact that DNA based trees are often more accurate for closely related part of tree; whereas protein based trees are more accurate for more long distances relationships. Further, a group of algorithms used in tree building may, in fact, out perform others under certain scenarios. The codon alignments for each ortholog gene cluster and the corresponding ML trees were then used as input for the ML codon substitution analysis, which will be described in the next section. The construction of phylogenetic trees depended on maximization of a likelihood function, but up to now no details as to how such a likelihood function was constructed have been given. To gain more insight into the ideas used to construct such likelihood functions, it is suggested that a reader consult chapter 4 of the book by Yang (2006).

Presented in Figure 14.7.1 is a unrooted fourteen species phylogenetic tree used in the maximum likelihood (ML) analysis. An asterisk "$*$" denotes species with high quality genome sequences to which linage specific branch sites tests were applied. This tree is also the consensus tree for the phylogenetic reconstruction. Species images were republished with permission from the Ensembl project - Sanger Institute, see Raj (2009) for more details.

14.8 Overview of Markovian Codon Substitution Models and Their Applications in Comparative Mammalian Genomics

A parameter denoted by ω was used in a Markovian codon substitution process and is defined as follows. For a protein coding gene let d_N denote the number of non-synonymous substitutions (amino acid altering) per nonsynonymous site and let d_S denote the corresponding number of synonymous (silent mutations) per synonymous substitution. Then, ω is defined as the ratio $\omega = d_N/d_S$ and provides a measure of selection pressure at the amino acid sequence level. If the phrase "d_N d_S ratio" is entered into a search engine for the internet, then many sites will appear, and some of these sites will be devoted to a description of how to estimate this ratio by comparing two strands of DNA in which one is ancestral and one is derived. A value of $\omega > 1$ is interpreted as evidence for positive selection; whereas a value such that $\omega < 1$ is interpreted as the gene was under negative selection pressure in the sense that frequency of the gene would decrease over time.

The Markovian codon substitution model used by Raj (2009) had the following ingredients. The state space $\mathfrak{S}$ of the model consisted of the 61 sense codons of the universal code and let $\boldsymbol{Q} = (q_{ij})$ denote the 61×61 rate matrix of the Markov process, where q_{ij} is the rate of substitution from sense codon i to sense codon j for $i \neq j$. Let β denote a scale, let κ denote that transition/transversion ratio and let π_j denote the equilibrium frequency of codon j as estimated from the observed data on codons. Values of the scale parameter $\beta > 0$ depend on the time scale under consideration.

For example, if time is measured in years, then $1/\beta$, which is proportional to the expected times spent by a Markov in some state, is expressed in years.

Given the above parameters, the elements of the rate matrix off the principal diagonal were defined as follows: $q_{ij} = 0$ if i and j differ at more than one site. If the substitution $i \to j$ was a synonymous transversion, then $q_{ij} = \beta\pi_j$, and if the substitution were synonymous transition, then $q_{ij} = \beta\kappa\pi_j$. On the other hand, if the substitution $i \to j$ were a nonsynonymous transversion, then $q_{ij} = \beta\omega\pi_j$, but if the substitution were nonsynonymous transition, then $q_{ij} = \beta\omega\kappa\pi_j$. Further discussions of the model under consideration may be found in Goldman and Yang (1994) and Yang (2006).

The rate matrix $\mathbf{Q}$ just described is that for a Markov jump process in continuous time, but in the applications of this model to genomic data organized into species and gene trees, the actual periods of time involved in the evolution of species is not, in general, known. Yet this model formed a basis for calculating likelihood functions. From the discussion of this model it is not clear as to how the likelihood functions were calculated. However, it is worthwhile to make an informed guess. In chapter 6, in which the mathematical structure underlying Markov jump processes in continuous time was discussed, a technique was described for deriving the one step transition matrix for the discrete time Markov chain embedded in a Markov jump process in continuous time from the rate matrix $\boldsymbol{Q}$. Briefly, this technique consisted of normalizing each row the matrix $\boldsymbol{Q}$ to produce a Markov transition matrix $\boldsymbol{P}$ such that the all rows summed to one. If this idea was indeed used, it would have been a straight forward exercise to write down the likelihood for a set of codon transitions, but when two sequences of codons are compared, an investigator would need to assume which sequence was ancestral and which one was derived.

Alternatively, if an investigator had some idea as the time two or more species under consideration had a common ancestor, then this time could be used as the origin for calculations based on a Markov jump process in continuous time. For this case however, the matrix $\boldsymbol{Q}$ would need to be modified by specifying its diagonal elements. Let $q_i = q_{ii}$ denote the diagonal elements of a rate matrix $\boldsymbol{Q}_1$, which is constructed from the matrix $\boldsymbol{Q}$. Then for state $i \in \mathfrak{S}$, the diagonal elements of the matrix $\boldsymbol{Q}_1$ would have the form $q_i = -\sum_{j \neq i} q_{ij}$ and the off-diagonal elements of $\boldsymbol{Q}_1$ would correspond to those in the matrix $\boldsymbol{Q}$, which were defined above. For this continuous time model, the matrix of transition probabilities is $\boldsymbol{P}(t) = (p_{ij}(t)) = \exp(\boldsymbol{Q}_1 t)$, the exponential matrix. Given the elements of this matrix, a likelihood function could be written down in terms of the elements of the matrix $\boldsymbol{P}(t)$ using the Markov property in continuous time.

The substitution process just described was used in the context of branch-site models in which the ratio ω varies among the codons under observation. In particular, the model used in the analysis of the genomic data from the mammalian species studied had four site classes such that each codon in a sequence belonged to one of the four classes. The frequencies of site classes were characterized in terms of two parameters $p_0 > 0$ and $p_1 > 0$ such that $0 < p_0 + p_1 \leq 1$. Class 0 contained all codons that were conserved in all linages and the parameters p_0 and ω_0 were estimated under the constraints $0 \leq p_0 < 1$ and $\omega_0 \in (0, 1)$ by the method of maximum likelihood. This estimation procedure differed from that of Yang and Nielsen (2002) in which the parameter ω_0 was assigned the value $\omega_0 = 0$. In a second class of codons denoted by 1, which were thought to be weakly constrained by selection, the parameter $\omega = \omega_1$ was set equal to $\omega_1 = 1$, indicating this site was considered neutral. It was deemed advantageous to fix $\omega_1 = 1$ for weakly constrained sites so that they belonged to the set of neutral sites rather than being falsely assigned to sites undergoing positive selection. In the third and fourth classes, denoted by 2 and 3, in the background lineages the parameters ω_0 and ω_1 are either estimated or assigned values as in classes 0 and 1 but in the foreground branches the parameter ω_2 was required to be > 1 and was estimated from the data under this constraint.

In the foregoing discussion, it has been tacitly assumed that the branches on the phylogenetic tree are partitioned a priori into foreground and background lineages. In the foreground linages it is assumed that positive selection may occur with $\omega_2 > 1$, but all other branches of the tree represent the background lineages, where the sites are allowed to evolve under negative or purifying selection with $0 < \omega_0 > 1$ or without selection, *i.e.* the neutral case, such that $\omega_1 = 1$. The ideas just described are summarized in symbolic form in the table 14.8.1. It should be mentioned at this point in the discussion that the use of the word, site, in what follows refers to three letter codons.

The next step in the description of the data analysis based on branch models is that of providing an overview of the procedure for calculating log-likelihood scores. Let $n \geq 1$ denote the number of sites under consideration and let $\boldsymbol{X}_h$ denote a vector of codons at site $h = 1, 2, \ldots, n$ across all sequences in an alignment. Let the scalar valued random variable $\mathbf{Y}_h$, taking values in the set $(0, 1, 2, 3)$, denote the site class to which site h

Table 14.8.1 Parameters and Site Classes for the Branch Models

Site Class	Frequency	Background	Foreground
0	p_0	$0 < \omega_0 < 1$	$0 < \omega_0 < 1$
1	p_1	$\omega_1 = 1$	$\omega_1 = 1$
2	$(1 - p_0 - p_1) \frac{p_0}{p_0+p_1}$	$0 < \omega_0 < 1$	$\omega_2 > 1$
3	$(1 - p_0 - p_1) \frac{p_1}{p_0+p_1}$	$\omega_1 = 1$	$\omega_2 > 1$

belongs, and let $f(\boldsymbol{X}_h \mid \mathbf{Y}_h)$ denote the conditional probability of observing the vector $\boldsymbol{X}_h$, given a realization of the random variable $\mathbf{Y}_h$. If $\mathbf{Y}_h = 0$ or $\mathbf{Y}_h = 1$, all branches on the phylogenetic tree have the same ω ratio so that $f(\boldsymbol{X}_h \mid \mathbf{Y}_h)$ may be calculated following procedures used by Goldman and Yang (1994). But if $\mathbf{Y}_h = 2$ or $\mathbf{Y}_h = 3$, then the ω ratios differ for background and foreground branches so that in this case $f(\boldsymbol{X}_h \mid \mathbf{Y}_h)$ may be estimated using the procedure of Yang (1998).

Given these estimates, the estimate of the unconditional probability $f(\boldsymbol{X}_h)$ of observing vector $\boldsymbol{X}_h$ is given by the equation

$$f(\boldsymbol{X}_h) = \sum_{k=0}^{3} p_k f(\boldsymbol{X}_h \mid \mathbf{Y}_h = k). \tag{14.8.1}$$

Then, under the assumption that the substitution processes are independent among sites, it follows that the log-likelihood function is given by

$$\ell = \sum_{h=1}^{n} \ln f(\boldsymbol{X}_h). \tag{14.8.2}$$

The parameters of this function include branch lengths of the phylogeny, the transition-transversion rate ratio κ and the parameters in κ distribution. Given these parameters, they were estimated by a numerical maximization procedure, see for example Yang (1997).

Likelihood ratio tests were then constructed, following a procedure described by Yang (1998), to test two alternative hypotheses. For an alternative hypothesis denoted by H_a, it was assumed that $\omega_2 \geq 1$ so that positive selection did occur with respect to the gene under consideration. Under hypothesis H_a, the foreground lineage was allowed to have a subset of sites denoted by P_s with accelerated substitution rates denoted by $\omega_2 > 1$. Let S denote the set of codon sites under consideration. Then, in symbols $P_s = (s \in S \mid \omega_2(s) > 1)$. On the other hand, under the null hypothesis denoted by H_0, it was assumed that ω_2 was fixed at $\omega_2 = 1$, indicating that, by assumption, there was no positive selection acting on the gene in question. Under the null hypothesis, it was allowed that there were two classes of sites in the branches. In one class of sites denoted by P_1 it was assumed that $\omega_1 = 1$, indicating neutral evolution. In symbols, $P_1 = (s \in S \mid \omega_1(s) = 1)$. In the other set of sites denoted by P_0, it was assumed that $0 < \omega_0 < 1$, indicating these sites were "constrained" by selection. In symbols, $P_0 = (s \in S \mid 0 < \omega_0 < 1)$.

A likelihood ratio test (LRT) of the null hypothesis H_0 against its alternative H_a was carried out for each gene under consideration. A brief definition of this test is as follows. Let ℓ_0 denote the maximum of the log-likelihood function under the null hypothesis H_0. Similarly, let ℓ_a denote the maximum of the log-likely hood function under the alternative hypothesis H_a. Then, under regularity conditions, it is shown in books on mathematical statistics that the random variable

$$-2\left(\ell_0 - \ell_a\right) \tag{14.8.3}$$

is approximately distributed as a Chi-square random variable in large samples with degrees of freedom equal to $m - r$, where m is the dimension of the parameter space under the null hypothesis H_a and $r < m$ is the dimension of the parameter space under H_0. Thus, by calculating the statistic in (14.8.3) and using the distribution function of the Chi-Square distribution corresponding to the appropriate number of degrees of freedom, it was possible to compute a p value for each gene under consideration. All those genes that had statistically significant p values were considered possible candidates for the action of positive selection and were then tested for further confirmatory evidence in subsequent tests. With regard to caveats concerning the use of the Chi-square distribution approximation in these tests, the papers cited in Raj (2009) in this connection, may be consulted.

Among the $LRTs$ that were statistically significant, which provided evidence for positive selection, an empirical Bayes approach was used to estimate the posterior probability that a site belonged to a given class of sites. Let $f\left(\boldsymbol{Y}_h = k \mid \boldsymbol{X}_h\right)$ denote the posterior conditional probability that site h belongs to class k, given the vector of data $\boldsymbol{X}_h$ across an alignment. Then, by definition, this probability is given by the formula

$$f\left(\boldsymbol{Y}_h = k \mid \boldsymbol{X}_h\right) = \frac{p_k f\left(\boldsymbol{X}_h \mid \mathbf{Y}_h = k\right)}{f\left(\boldsymbol{X}_h\right)} = \frac{p_k f\left(\boldsymbol{X}_h \mid \mathbf{Y}_h = k\right)}{\sum_{k=0}^{3} p_k f\left(\boldsymbol{X}_h \mid \mathbf{Y}_h = k\right)} \tag{14.8.4}$$

for $k = 01, 2, 3$. The sites h with high posterior probabilities, $f(\boldsymbol{Y}_h = k \mid \boldsymbol{X}_h) > 0.95$, were inferred to be under positive selection so that this Bayesian approach made it possible to identify sites that were under selection even if the average ω ratio was less than one. The implementation of formula (14.8.4) requires maximum likelihood estimates of all parameters, which are subject to variation among samples. To take into account this uncertainty, a Bayes empirical Bayes approach due to Deeley and Lindley (1981) was used following an implementation of this approach by Yang *et al.* (2005).

Many $LRTs$ were performed on the same data so that a correction for multiple testing was required. Unless otherwise stated p values derived from $LRTs$ were based on False Discovery Rate (FDR) adjusted for multiple testing using the method of Benjamini and Hochberg (1995). The Mann-Whitney U test (MWU) was also used to test for statistically significant differences in overall selection pressure between Gene Ontology (GO) biological processes.

The MWU test was used to compare different groups of genes and for assessing whether two independent samples of p values were from the same distribution. The MWU test takes a list of p values from $LRTs$ and inspects it to determine if a specified category of p values tends to occur at the beginning or the end of a list. This classification based on MWU tests identifies categories of genes with small p values from $LRTs$, see Nielsen *et al.* (2005) for details.

It has been widely recognized that statistical evidence for positive selection of some gene would be considerably enhanced by additional evidence as to how a gene functions. Such evidence should be validated by experimental and functional assays, which would include establishing a direct link between observed amino acid changes, in post-translational modification and or efficiencies in catalyzing chemical process involved in the function or functions of a gene.

Like statistical procedures cited above, the search as to how a gene functions was greatly expedited by software packages which aid in searching data bases, but even a brief description of these packages will not be attempted here. In the remainder of this section, however, to a brief discussion of the biological processes involved in gene function and an overview of the massive number of calculations that were done by Raj (2009) in order to reach his goals.

As part of the process assessing evidence for positive selection on a gene, $3D$ protein structures were downloaded from a data base and visualization tools were used to study genes subject to positive selection by studying protein structures resulting from mutations in codons. Another resource that was used was the web site of The Gene Ontology project, which is a major bioinformatics initiative with the aim of standardizing the representation of genes and gene product attributes across species and databases.

In this connection, another resource was used called *DAVID*, Database for Annotation, Visualization and Integrated Discovery, which is a functional annotation clustering tool that was adapted to extract biological features of genes. The features included classification, biological pathways, disease associations, protein-protein interactions, protein functional domains, gene functional summaries, sequence features and gene expression.

The maximum likelihood estimation of parameters was carried out using a program called *codeML* within the *PAML*, Phylogenetic Analysis by Maximum Likelihood, package version 4.0 Yang, 1997, 2007. To carry out the massive among computing involved in maximum likelihood estimation, a net work of computers called *CamGrid* was used at University of Cambridge in England. To run these experiments, *codeML* was recompiled to run on *CamGrid*. For each gene and for each species, *codeML* was run under three test models, a branch model, a two-branch model and a null model. To ensure that in each case convergence to the best likelihood occurred, all *codeML* analyses were preformed twice. The computing time for 102,704 alignments of nine mammals was 15 days and, on average, 300 core processors were used concurrently. Altogether a total of 616,224 maximum likelihood iterations were run.

Among the many conclusion of these investigations was that there was a significantly smaller proportion of genes that were affected by positive selection in primates than in rodents and domestic animals. Within primates, the human linage had the smallest number of positively selected genes, which was attributed to the reduced efficacy of natural selection in humans due to their smaller long-term effective population size. The evidence indicated that more genes had undergone positive selection in domestic animals, dog, cattle and horse, than in other mammals, which suggested that it may be possible to detect more evidence for positive selection when longer branches of the evolutionary tree are tested.

It is thought by many that domestic animals have undergone a transition from wild to domestic form within the last 10,000 to 15,000 years. There is also convincing evidence that domestic plants such as wheat, barley, oats and field corn, have undergone rather rapid evolution during this time period. The case of field corn, Zea maize, is particularly interesting and it is suggested that a reader consult the internet for more information. It has long been thought in some circles that humans have evolved very little during the past 10,000 during the process of organizing into villages, cities and nations. For a contrary view of human evolution during the last 10,000 or so years, it is suggested that a reader consult the interesting book by Cochran and Harpending (2009). Among the many topics included in this book is evidence that genes for resistance to malaria, blue eyes and lactose tolerance have evolved in humans during the last 10,000 years. Interestingly, the computer experiments reported in chapters 11 and 12 suggest that it is indeed the case that new beneficial mutations may arise and become established in populations during periods of 10,000 to 15,000 years.

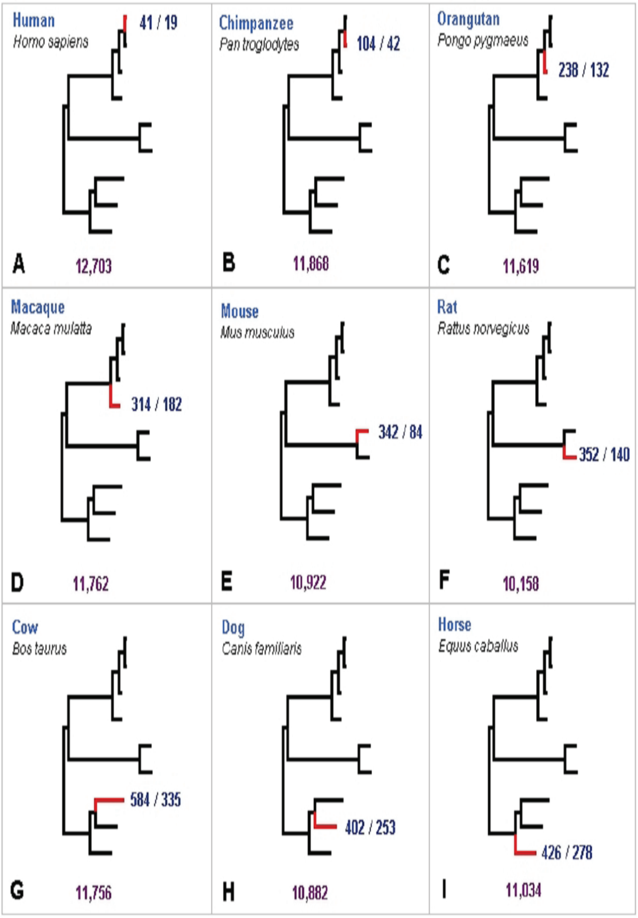

Figure 14.8.1 Genes Under Positive Selection in Nine Mammalian Genomes.

Figure 14.8.1 contains a graphical summary of the results obtained by Raj (2009). Panels *A* through *I* show the lineage specific tests for positive selection with the foreground branches in red. The numbers next to each terminal branch in each panel represent the number of positively selected genes identified by each LRT such that $p < 0.05$ followed by the number of $PSGs$ (Positively Selected Genes) after applying a multiple correction ($FDR < 0.05$) based on the Benjamini and Hochberg (1995) method. For example, for humans in panel A these numbers are 41 and 19, respectively. The total number of orthologous gene sets tested for each species is listed at the bottom of each panel.

In the remaining sections of this chapter, preliminary stochastic models for accommodating genetic recombination as well as various types of mutations such as nucleotide substitutions, deletions, insertions, duplications and inversions will be developed. These preliminary developments are in a sense a return to the issues discussed the closing sections of chapter 8 in which mutation other than nucleotide substitutions were discussed.

14.9 Probabilistic Methods for Simulating Genetic Recombination at the Molecular Level

At the Mendelian level, linkage between two loci on the same chromosome is expressed in terms of a recombination probability ρ, which was discussed in detail in chapter 2. There is also a notion of genetic distance between two loci, which is expressed in terms of a unit called a centimorgan. This unit was named in honor of T. H. Morgan, one of the early influential workers in Drosophila genetics. A reader, who is interested in more details, may wish to consult the internet for information or read one of the text book on genetic cited in previous chapters, where information on early workers in genetics may be obtained. One centimorgan, for example, corresponds to a recombination probability with the value $\rho = 0.01$, which may be thought of as rather "tight" linkage between two loci such that the genetic distance between the two loci would be small. Whereas, a value of ρ such that $\rho = 0.25$ would be interpreted as a greater genetic distance between two loci. Genetic map functions are concerned with relating the recombination probability to the actual physical distance d between two loci measured on some scale. No discussion of these functions will be included in this section, but if a reader is interested in more information and a review of literature on genetic map functions and their applications in linkage studies at the Mendelian level, the paper by Zhao and Speed (1996) may be consulted. The reason why these functions will not be discussed here is that at the molecular level, the distances among markers, such as $SNPs$, may be measured in terms of the number of bases between any two of them when a strand of DNA is viewed in a linear form, and, thus, in such cases a measure of the physical distance between two markers is known.

When an investigator is interested in constructing a model of a genomic region with a view toward using this model and Monte Carlo simulation methods to study the evolution of this region in a population over a large number of generations in which mutation and genetic recombination occur, a question that naturally arises is: how many base pairs, *bps*, correspond to one centimorgan? The answer to this question is that it depends on the species under consideration. In humans, for example, by consulting many sites on the internet, it can be seen there seems to be a consensus that on average one centimorgan corresponds to one million base pairs or 1 Mb. If, for example, the computer memory available to an investigator is limited to considering genomic regions of 1 Mb, then, if the evolution of a human genomic region is under consideration, attention would have to confined to values of about $\rho = 0.01$ for the recombination probability between two markers. Of course, since this number is an average, an investigator would be free to consider larger values of ρ in his computer simulation experiments even if a human genomic region were under consideration.

The human malaria parasite, Plasmodium falciparum, which is responsible for hundreds of millions of cases of malaria each year and kills more than one million African children annually, is in a haploid phase in humans, but in mosquitoes it has a diploid phase so that genetic recombination may occur. Gardner *et al.* (2002) reported on an analysis of the genome of this parasite and found that it had 23 megabase nuclear genome that consists of 14 chromosomes, encodes about 5,300 genes and is the most $(A - T)$ rich genome that had been sequenced up the date the research was completed. Due to the diploid phase of the parasite in mosquitoes, it is possible to cross strains and study recombination in its genome as well as types of mutations. Such a study was carried out Su *et al.* (1999) to produce a high-resolution genetic map of the genome of this parasite. This team of investigators used a genetic cross to construct a map of 901 genetic markers that fell into 14 linkage groups corresponding to 14 chromosomes of the species. Meiotic crossover activity was observed to be high and it was estimated that one centimorgan corresponded to about 17 kilobases and was notably uniform over each chromosome length. Gene conversion events and spontaneous changes in lengths of microsatellites were also evident in the inheritance data. For those readers who are interested in more details regarding this research, the paper of Su *et al.* (1999) cited above should be consulted.

With regard to simulating the evolution of a genomic region, the processing of simulated data based on a centimorgan corresponding to about 17,000 base pairs would be easier in terms of the number of bases that would need to be considered in a model genomic region than that for the case in which a centimorgan corresponded to 1 Mb. When a network of computers is available to an investigator, it may be possible to consider model genomic regions with a sufficiently large number of base pairs such the 1 centimorgan would correspond to 1 Mb, but before embarking on such an experiment, estimates of the computer time required to complete an experiment would be needed. Such estimates would be particularly crucial when one considers the size of a model genomic region that would be required when it was desired to simulate genetic recombination with respect to three or more markers on a chromosome over a large number of generations. Another problem that must be confronted when planning such experiments would be that of parameterizing the linkage distribution when genetic recombination is to be studied with respect to two or more markers on a chromosome.

Before launching into a discussion of genetic recombination, it will be helpful to know the phase of meiosis in which chiasmata or crossing over occurs. Meiosis consists of two cell divisions, Meiosis I and II. A description of Meiosis II will be given subsequently, but it is important to know that in the Prophase I of meiosis I, DNA replication occurs, and it is during this replication that such mutations as nucleotide substitutions, deletions and insertions may occur. In subsequent sections, models of such mutations that occur on a per generation time scale will be introduced. The replication of DNA in a diploid genome results in the doubling of the amount of DNA in a cell that will eventually undergo meiosis. Thus, if c denotes the amount of haploid DNA in a genome, then the amount of DNA in a diploid cell is initially $2c$ and after DNA replication it is $4c$. During this $4c$ phase, homologous chromosome pairs are formed along a spindle such that one chromosome was contributed by the female parent and one from the male parent. But the assembly of maternal and paternal chromosomes along the spindle appears at random so that if a diploid species has n pairs of chromosomes, then the total number of ways maternal and paternal chromosomes may be distributed along the spindle is 2^n. Moreover, if the distribution of parental chromosomes along the spindle is random, then each configuration occurs with probability $1/2^n$. In humans, $n = 23$ and it is thought by many that this random distribution of maternal and paternal chromosomes along the spindle during meiosis is a significant factor contributing to the high level of individuality that is observed among humans within and among families.

Each pair of chromosomes consists of two chromatids and within each chromosome, the chromatids are referred to as sister chromatids, but the chromatids contributed by the female and males parents are referred to as non-sister chromatids. Altogether, each pair of homologous chromosomes is made up of 4 chromatids. The process of crossing over, resulting in genetic recombination of maternal and paternal DNA, occurs among non-sister chromatids. The crossing over process is reciprocal in the sense that if a segment of maternal DNA is replaced by the segment of paternal DNA, the corresponding segment of maternal DNA is inserted into the paternal DNA. It is also important to recognize that during this phase of meiosis, the process of gene conversion may occur, which arises as a result on non-reciprocal or unequal crossing over. In meiosis I, a $4c$ cell divides and gives rise to two $2c$ cells. In meiosis II there is no replication of DNA and each of these $2c$ cells divides into two cells which gives rise to 4 cells. Each these cells, gametes, contains only the haploid amount c of DNA, which is characteristic for a particular species. With respect to each chromosome, it is thought that the haploid content of the DNA in the four gametes corresponds to that in the four chromatids that existed after crossing over occurred. In the next section, the modelling of the process of gene conversion will be considered.

Having sketched the process of meiosis, we will next turn to the problem of simulating genetic recombination in a model of a autosomal genomic region denoted by G. It will be assumed that G consists of two complementary strands of DNA. Let d denote the number of nucleotides in each strand in G, and let M_1 and M_2 denote two $SNPs$ which will serve as markers for the study of genetic recombination with respect to two loci at the molecular level. The number of nucleotides between M_1 and M_2 will be denoted by d_{12}, and, by design, this number will be chosen such that it is at least one centimorgan for the species under consideration. Let m_1 and m_2 denote the alleles of M_1 and M_2, respectively. One approach to simulating genetic recombination among the four alleles under consideration would be that of formulating a model for simulating the number of chiasmata in the region between the markers M_1 and M_2 and observe that if there is an odd number of chiasmata in this region, then a recombination with respect to M_1 and M_2 would occur, but if the number of chiasmata was even, then no recombination would be observed. Rather than taking this route, the procedures for modelling recombination developed in chapter 2 will also be used in this section. The reason for taking this route is that when simulating the evolution of recombination in a model genomic region, it will be possible to detect genetic recombination only with respect to a set of markers that could be identified easily in a computer.

Just as in section 2.4 in which the case of two linked loci was considered, let $(00, 11)$ denote an arbitrary genotype with respect to two autosomal loci such that 00 represents the alleles contributed by the maternal parent and 11 represents the alleles contributed by the male parent. With respect to alleles under consideration the symbol 00 may represent any of the four combination alleles present at the two loci; namely, M_1M_2, M_1m_2, m_1M_2 and m_1m_2. Similarly, the symbol 11 may also represent any of the four combinations of alleles just listed. The set of gametes that an arbitrary genotype $(00, 11)$ may produce will be denoted by $(00, 10, 01, 11)$. Observe that 01 and 10 are recombinant gametes; whereas 00 and 11 are nonrecombinant gametes, because they are copies of the alleles in the maternal and paternal parents, respectively. Let $(\gamma(00), \gamma(10), \gamma(01), \gamma(11))$ denote the set of probabilities corresponding to the four types of gametes. As regular crossing over is a reciprocal process, there exists a set of equalities for these probabilities. That is $\gamma(00) = \gamma(11)$ and $\gamma(01) = \gamma(10)$. Therefore, to express these probabilities as a function of a recombination probability ρ for the two loci, it suffices to consider the column vector

$$\boldsymbol{\gamma}_1 = \begin{pmatrix} \gamma(00) \\ \gamma(10) \end{pmatrix}. \tag{14.9.1}$$

As was shown in section 2.4, the equation $\lambda = 1 - 2\rho$ plays an important role in developing a structure that may be generalized to three or more linked loci on the same chromosome. Let $\boldsymbol{\lambda}$ denote a column vector defined by

$$\boldsymbol{\lambda} = \begin{pmatrix} 1 \\ \lambda \end{pmatrix}, \tag{14.9.2}$$

and let

$$\boldsymbol{A}_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \tag{14.9.3}$$

denote a 2×2 symmetric matrix. Then, the vector-matrix equation connecting the recombination probability ρ with the vector $\boldsymbol{\gamma}_1$ is

$$\boldsymbol{\gamma}_1 = \frac{1}{2^2}\boldsymbol{A}_2\boldsymbol{\lambda}. \tag{14.9.4}$$

At this point, a reader may wish consult section 2.4 for more details. From this equation, it can be seen that by assigning a numerical value to the parameter ρ in the interval $\left[0, \frac{1}{2}\right]$, the vector $\boldsymbol{\gamma}_1$ would be completely determined numerically.

To simulate the gametic output of the process of meiosis, it will be necessary to define a 4×1 vector $\boldsymbol{\gamma}$ of gametic probabilities as function of the recombination probability ρ. As $\gamma(00) = \gamma(11)$ and $\gamma(10) = \gamma(01)$, it follows that

$$\boldsymbol{\gamma} = \begin{pmatrix} \gamma(00) \\ \gamma(10) \\ \gamma(01) \\ \gamma(11) \end{pmatrix} = \begin{pmatrix} \gamma(00) \\ \gamma(10) \\ \gamma(10) \\ \gamma(00) \end{pmatrix}. \tag{14.9.5}$$

As an illustrative example as to how this gametic distribution may be applied when simulating the evolution of a diploid population, for the case to two linked markers each with two alleles as discussed above, consider the mating

$$(M_1M_2, m_1m_2) \otimes (M_1M_2, m_1m_2), \tag{14.9.6}$$

where the genotype of the female is on the left and that of the male on the right.

Observe that in this mating, both the female and male are heterozygous at both loci. Moreover, the alleles contributed by the maternal parent of the female are M_1M_2, and similarly the alleles contributed by the female's paternal parent are m_1m_2. The same remarks hold for the male on the right in (14.9.6). Therefore, the set of possible gametes for the female per meiosis is $(M_1M_2, M_1m_2, m_1M_2, m_1m_2)$ which occur with the corresponding probabilities in the vector $(\gamma(00), \gamma(01), \gamma(10), \gamma(11)) = \boldsymbol{\gamma}^T$, where T stands for transpose of a vector or matrix. If it is assumed that the male produces the same set of gametes with the same probabilities per meiosis, then, when the contributions of the female and male parents are represented in the genotypes of the set of possible offspring of the mating in (14.9.6), there are 16 possible genotypes for the offspring. In the presence of linkage, the 16 probabilities corresponding to each of these genotypes may be represented as the elements in the 4×4 matrix

$$\boldsymbol{O}_p = \boldsymbol{\gamma}\boldsymbol{\gamma}^T = (P\left[(x_1x_2, y_1y_2)\right]), \tag{14.9.7}$$

where (x_1x_2, y_1y_2) is an arbitrary genotype among the 16 possible genotypes of the offspring from the mating in (14.9.6), and, for the moment, the sex of an offspring is ignored.

To simulate the genotype per offspring from a mating in (14.9.6), one would need to consider a vector $\boldsymbol{p}$ of multinomial probabilities of 16 dimensions, which would be constructed from the matrix $\boldsymbol{O}_p$ by setting $\boldsymbol{p} =, \boldsymbol{O}_p$, where the symbol, indicates the first four elements of $\boldsymbol{p}$ are the first row of $\boldsymbol{O}_p$, the next four elements are those in the second row and so on until a row vector of 16 dimensions is constructed. Then, the distribution of the genotype of an offspring from the mating in (14.9.6) has a multinomial distribution with sample size or index 1 and probability vector $\boldsymbol{p}$. If the mating contributes more than one offspring to the population, then the total number of offspring produced by mating could be simulated by following the procedures for simulation realizations from a multinomial distribution with some index $N \geq 1$ and probability vector $\boldsymbol{p}$ as described in chapters 11 and 12. It is also essential to mention that the procedure just described for the mating type in (14.9.6) would have to be carried out for mating types in the set of all mating types. For illustrations of this set of possible matings for the case of one autosomal locus with two alleles as well as various mating systems, a reader may consult chapters 11 and 12 for details. On the other hand, if attention were confined to random mating, then the ideas developed in chapter 3 could be implement in models in which the availability of computer memory would determine the number of loci and the number of alleles per locus that could be considered in a computer experiment.

For the case of two linked loci with two alleles at each locus under consideration, when the contributions of the gametes contributed by the female and male parent are taken into account, there would be $16 \times 16 = 256$ types of mating to consider, but we will not go into any further details here except to state that if further simplifying assumptions were introduced, such as representing genotypes only in terms of phenotypes, the number of mating types to be considered could be significantly reduced. When considering this set of mating types, when writing software to implement the production of offspring by genotype, it is advisable to use a standard ordering of the four types of gametes to be considered in a model with two loci with two alleles at each locus. One such standard ordering would be that in the set $(M_1M_2, M_1m_2, m_1M_2, m_1m_2)$ for both the female and male in a mating. It should also be noted the ordering of the set of gametic probabilities $(\gamma(00), \gamma(01), \gamma(10), \gamma(11))$ would depend on the genotypes of the female and male of a mating. In the next section, this point will be discussed in more detail.

It should also be mentioned that if it were desired, three linked markers on an autosomal chromosome could also be considered using the concepts developed in section 2.5, but an extension of the procedures developed in this section to the case of three linked markers will be left as an exercise for an interested reader. It also needs to be mentioned that the introduction of the parameter $\lambda = 1 - 2\rho$ for the case of two loci, greatly expedites the extension of the model to the case of three linked loci.

14.10 A Preliminary Probabilistic Model of the Process of Gene Conversion

In this section, a preliminary model of gene conversion will be formulated. As a starting point in formulating the process of gene conversion, suppose the genotype of an individual is

$$(M_1 M_2, m_1 m_2). \tag{14.10.1}$$

As illustrated in section 14.3, the process of gene conversion is the result of unequal or non-reciprocal crossing over during the during the process of meiosis, and, as was mentioned in section 14.9, crossing over occurs that phase of meiosis in which pairs of maternal and paternal chromatids are aligned along the spindle of dividing quadra-ploid cell. It will also be assumed that no mutations, such as nucleotide substitutions, occur during this phase, but they will assumed to occur during the cell phase when copies of DNA are produced as will be described in the next section. With the exception of such organisms as yeast, very little is known about the process of gene conversion in other species. For an account of detecting gene conversion in yeast and a graphic description of the gene conversion process, it is suggested that the reader consult Figure 1.3 in W. H. Li (1997). In the approach to formulating a stochastic model of the gene conversion process described in this section, conversion events will be distinguished as to whether an event involved maternal or paternal DNA. However, it should be stated that this approach may be an over simplification of the process, but, it does, however, seem to be a good starting point for a systematic formal mathematical description of the process, which can easily be modified in subsequent research.

A step towards representing this process of genetic recombination during meiosis symbolically for the case of the genotype in (14.10.1) at the level of a quadra-ploid cell, consider the 4 × 5 table

$$\begin{array}{|c|c|c|c|c|}
\hline
5' & M_1 & E & M_2 & 3' \\
\hline
3' & M_1' & E' & M_2' & 5' \\
\hline
3' & m_1' & e' & m_2' & 5' \\
\hline
5' & m_1 & e & m_2 & 3' \\
\hline
\end{array} . \tag{14.10.2}$$

All letters in this table represent strands of DNA, and the first two rows of the table represent the two chromatids contributed by the maternal parent and the lower two rows represent the two chromatids contributed by the paternal parent. Symbols with a prime represent copies of strands of DNA that arose in the doubling phase to produce a quadra-ploid cell. As indicated in the table the marker M_1 and its allele m_1 occur in regions of DNA located in the second column of the table, and, similarly, the marker M_2 and its alleles m_2 occur in regions of DNA in the forth column of the table.

To symbolize an example of the process of gene conversion, consider that 4 × 5 table

$5^{\prime}$	M_1	E	M_2	$3^{\prime}$
$3^{\prime}$	$M_1^{\prime}$	$E^{\prime}$	$M_2^{\prime}$	$5^{\prime}$
$3^{\prime}$	$m_1^{\prime}$	$e^{\prime}$	$M_2^{\prime}$	$5^{\prime}$
$5^{\prime}$	m_1	e	m_2	$3^{\prime}$

, (14.10.3)

which provides a symbolic example of a non-reciprocal exchange of DNA between two non-sister chromatids. As represented in row three of the table the maternal DNA segment row 2 of the table containing the marker $M_2^{\prime}$ has been inserted into a paternal DNA segment that originally contained the marker $m_2^{\prime}$, but the paternal DNA segment containing the marker $m_2^{\prime}$ was not inserted into the segment of maternal DNA containing the marker $M_2^{\prime}$. That is the exchange of DNA segments was not reciprocal.

As an another example of the process of gene conversion consider the 4 × 5 table

$5^{\prime}$	M_1	E	M_2	$3^{\prime}$
$3^{\prime}$	$M_1^{\prime}$	$E^{\prime}$	$m_2^{\prime}$	$5^{\prime}$
$3^{\prime}$	$m_1^{\prime}$	$e^{\prime}$	$m_2^{\prime}$	$5^{\prime}$
$5^{\prime}$	m_1	e	m_2	$3^{\prime}$

. (14.10.4)

In this case, the paternal DNA segment in row three of the table containing the allele $m_2^{\prime}$ has been inserted into the maternal segment originally containing the marker $M_2^{\prime}$ in row 2 of the table but the reciprocal exchange of DNA did not occur. It should be noted that if the DNA represented in column 2 of the table (14.10.2), one could set down examples for the process of gene conversion occurring in the DNA segments containing the marker M_1 and its allele m_1. Thus, for the case of two loci, there would be four types of gene conversion.

If the process of meiosis and crossing over were normal, then, for the case of two loci, one would expect the following complete set of gametes

$$(M_1M_2, m_1M_2, M_1m_2, m_1m_2)$$

to be produced with respect to the two markers, and with the corresponding gametic distribution $((1-\rho)/2, \rho/2, \rho/2, (1-\rho)/2)$. To help fix ideas the order used in the listing of the complete set of possible gametes when normal crossing over occurs will also be used in all the cases enumerated below. When the process of gene conversion occurs however, then as will be shown by examples, the complete set of gametes will not be produced. In what follows, all probabilities will be conditional on the event that a gene conversion has occurred during meiosis. Let η denote the conditional probability per meiosis that a gene conversion event occurs. Then, $1-\eta$ is the probability per meiosis that normal crossing over between non-sister chromatids occurs.

Consider, for example, the conversion event represented in table (14.10.3) and suppose all probabilities are conditional on the event that a gene conversion occurs. If the rows in this table represent the set of four gametes that will be produced by meiosis, then, with respect to the two markers for loci under consideration, this set of gametes may be symbolized by $(M_1M_2, M_1^{'}M_2^{'}, m_1^{'}M_2^{'}, m_1m_2)$. However, from the genetic point of the gametes M_1M_2 and $M_1^{'}M_2^{'}$ are equivalent so that (M_1M_2, m_1M_2, m_1m_2) is the set of gametes produced by the process of gene conversion, and these gametes occur according to the ratios of $2:1:1$. Observe that in this case, the gamete M_1m_2 is missing. Consequently, under the assumption that the four rows are conditionally and uniformly distributed given the event that a gene conversion occurs during meiosis, it follows that the gametic distribution corresponding to the conversion event symbolized in (14.10,3) is $(2/4, 1/4, 0, 1/4)$). On the other hand, if the process of gene conversion represented in table (14.10.4) occurred, then (M_1M_2, M_1m_2, m_1m_2) is the set of gametes that would be produced according to the ratios of $1:1:2$. Note, in this case, the gamete m_1M_2 is missing. Therefore, the gametic distribution corresponding to this conversion event is $(1/4, 0, 1/4, 2/4)$.

By invoking the assumption stated above, when the process of gene conversion occurs in the DNA in the second column of table (14.10.1) representing locus 1 in the case analogous to table (14.10.3), then the set of gametes produced by meiosis would be (M_1M_2, M_1m_2, m_1m_2) in the ratios $2:1:1$. In this case, the gamete m_1M_2 is missing. Therefore, the gametic distribution corresponding to this conversion event is $(2/4, 0, 1/4, 1/4)$. If the process of gene conversion occurs at the first locus and is analogous to that in table (14.10.4) at locus 1, however, the set of gametes produced by meiosis would be (M_1M_2, m_1M_2, m_1m_2) in the ratios $1 : 1 : 2$ and in this case the gamete M_1m_2 is missing. Therefore, the gametic distribution corresponding to this event is $(1/4, 1/4, 0, 2/4)$.

With respect to the writing of software, it will be helpful to arrange the four gametic distributions just discussed in the form of a 5×5 table. In the first row of the table below, the four types of gametes are listed in the standard order as a basis for reference as to which gamete is missing. In the rows labeled 1 and 2, for example, the gametic distributions corresponding to the two gene conversion events that may occur with respect to locus 1, and in the two rows labeled 3 and 4 are the gametic distributions corresponding to the two gene conversion events that may occur with respect to locus 2.

$\cdot$	M_1M_2	m_1M_2	M_1m_2	m_1m_2
1	2/4	0	1/4	1/4
2	1/4	1/4	0	2/4
3	2/4	1/4	0	1/4
4	1/4	0	1/4	2/4

(14.10.5)

When the reproductive cell of an individual with the genotypes listed in (14.10.1) undergoes the process of meiosis and a gene conversion event occurs, then only one of the four gametic distributions listed in table (14.10.5) will be apply. Under the assumption that the four gene conversion events occur with uniform probabilities, given that a gene conversion even occurs, each of the four gametic distributions displayed in the table would be selected with probability 1/4 in a Monte Carlo simulation experiment, and the gametic distribution governing the production of gametes by this individual would be that the randomly selected gametic distribution.

On the other hand, suppose the genotype of an individual is

$$(M_1m_2, m_1M_2). \tag{14.10.6}$$

In the genotype presented in (14.10.1), the alleles are said to be in coupling phase; whereas, these in (14.10.6) are said to be repulsion phase. Consequently, for the case of regular crossing over, the gametes M_1m_2 and m_1M_2 are the parental or non-recombinant gametes and the gametes M_1M_2 and m_1m_2 are the recombinant gametes. Thus, for the case of regular crossing over the gametes during meiosis, the set $(M_1M_2, m_1M_2, M_1m_2, m_1m_2)$ would be assigned the probabilities $\rho/2, (1-\rho)/2, (1-\rho)/2, \rho/2$, respectively. However, if a gene conversion event did occur during meiosis, then one could, by using the line of reasoning outlined above, derive a set of gametic distributions analogous to those in (14.10.5), but the details of this exercise will be left to the reader. If one distinguishes maternal and paternal contribution to the genotype of an individual, then for the case of two loci each with two alleles, 16 genotypes could be distinguished, but the details will not gone into here except for mentioning three classes of cases.

One case, would be a genotype of the form (m_1M_2, M_1m_2) in which the maternal and paternal DNA are interchanged with respect to the genotype in (14.10.6). In this case, it would seem to be a straight forward exercise to modify the procedure used for the genotype in (14.10.6) to derive the four gametic distributions when the process of gene conversion occurs in the genotype (m_1M_2, M_1m_2). Another class of case would be those in which a genotype is homozygous at one locus but heterozygous at the other. An example of such a genotype would be (m_1m_2, M_1m_2). For this genotype gene conversion would be detectable at locus 2 so it would suffice to consider only gene conversion for locus 1 and how it affect the standard $1:1$ ratio of the alleles M_1 and m_1. Finally, for those genotypes that are homozygous at both loci, no detectable gene conversion events would be observed so that this class of genotypes only one type of gametic would be produced per meiosis.

We close this section with an illustration as to how the complementary events of normal crossing over or gene conversion would be carried out in a Monte Carlo simulation experiment involving the production of one offspring. For the case of a two sex population, consider an arbitrary mating

$$(x_1x_2, y_1y_2) \otimes \left(x_1^{'}x_2^{'}, y_1^{'}y_2^{'}\right), \tag{14.10.7}$$

where the genotype of the female is on the left and that of the male on the right. The first step is to compute a uniform random number U from the interval $[0, 1]$. If $U \leq \eta$, then a gene conversion event would be simulated following and a gametic distribution would be computed following the procedure developed in this section that would be appropriate for the mating in (14.10.7). On the other hand, if $U > \eta$, then the process of meiosis would be normal and a gametic distribution with respect to two linked loci developed in section 14.9 would be used to compute a gametic distribution.

Let $\boldsymbol{\gamma}_f$ denote the simulated gametic distribution for the female arranged in a 1×4 column vector, and let $\boldsymbol{\gamma}_m$ denote the simulated gametic distribution for the male, which would also be arranged in a 1×4 column vector. Then, consider the outer product

$$\boldsymbol{\gamma}_f \boldsymbol{\gamma}_m^T, \tag{14.10.8}$$

which is a 4×4 matrix containing 16 elements. Arrange these 16 elements into 1×16 row vector called $\boldsymbol{p}$. Then, the random genotype of the offspring would be distributed as a sample of size 1 from a multinomial distribution with the probability vector $\boldsymbol{p}$. If this mating were to produce some number N of offspring such that $N > 1$, then this simulation procedure just outlined would be repeated N times.

Of course, many details that would arise when simulating normal crossing over and gene conversion at two linked loci have be omitted, but the above description is a core and central concept for writing software for the Monte Carlo simulation of the two meiotic processes under consideration. In a research program designed to produce software for the two locus case, it is suggested that a team of investigators write software for simulating a normal meiosis and one in which gene conversion may occur for the case of one autosomal locus before attempting the case of two linked loci that was developed in this section.

14.11 Nucleotide Substitutions During Meiosis

It is thought that nucleotide substitutions occur at the time a molecule of DNA is replicated during cell division. Therefore, when any cell in the body divides, there is a risk that one or more nucleotide substitutions will occur somewhere in the genome. If such mutations occur in a somatic cell, there is no risk of it being passed on to the next generation, unless, during the development of an individual from a zygote, the mutation occurs in a linage of cell lines that develop into those organs that produce the gametes. In humans, these organs are the ovaries in females and the testicles in males. From the point of view of constructing a model of the process of nucleotide substitution that evolves on the time scale of generations, any attempt to take into account the occurrence of mutations in cell lines that develop into organs that produce the gametes for the next generation would be a very difficult task, because such lines are composed of millions of cells. Therefore, in this chapter the occurrence of nucleotide substitutions, will be confined to that phase of meiosis in which the DNA content of cell is doubled by the action of a DNA copying process. It will also be assumed that the replication of DNA is semi-conservative in the sense that during meiosis a double stranded DNA molecule splits and the resulting two single strands of DNA act as templates for the construction two doubled stranded DNA molecules that are formed by the binding of complementary bases.

In both these doubled stranded molecules of DNA, one strand consists of a parental strand and the other strand consists of bases that have been manufactured in the cell or, perhaps, manufactured in other cells and have been absorbed into the cell prior to the start of meiosis.

Before proceeding to describe the stochastic process that forms a mathematical basis for the biological process of nucleotide substitution during meiosis, it will be helpful to compare the formulation in this section with those in chapters 6,7 and 8. In these preceding chapters, the process of nucleotide substitution was viewed on an evolutionary time scale, which was usually expressed in years, and it was supposed that nucleotide substitutions occurred on this time scale without reference to the process of meiosis which would have occurred in each generation and within any set of species under consideration. On the other hand, in this section the stochastic model to be formulated, nucleotide substitution and other types of mutations will be viewed from the perspective of meiosis that occurs in connection with the generation gametes in every parental generation that produces the next generation of offspring. Thus, the process of mutation will be taken into account on time scale of generations rather than in evolutionary time measured in years.

Let GG denote an autosomal region of a genome in which a stochastic model of the process of nucleotide substitution is to be formulated, and suppose this genomic region consists of two strands of DNA with $d \geq 1$ base pairs. Symbolically, these two strands may be represented by the pair (s_1, s_2), where s_1 denotes a strand and s_2 is its complementary strand. During the replication of the DNA molecule denoted by the (s_1, s_2) two molecules of doubled stranded DNA will be produced, which will be denoted by the pairs $(s_1, s_1^{'})$ and $(s_2, s_2^{'})$. In the pair $(s_1, s_1^{'})$ the s_1 denotes the parental strand and $s_1^{'}$ it complementary strand of bases that has been constructed by a copying process using strand s_1 as a template. The molecule represented by the pair $(s_2, s_2^{'})$ has an analogous interpretation. From the point of view of constructing of stochastic model of the process of nucleotide substitution that occurs during the process of meiosis, it will be assumed that mutations may occur in the strands $s_1^{'}$ and $s_2^{'}$, each consisting of $d \geq 1$ bases.

In each of the strands s_1, $s_1^{'}$ and $s_2, s_2^{'}$, suppose the sites that the bases occupy are numbered in order from the left to the right from 1 to d. Then, this ordering may be denoted by the array $(1, 2, \ldots, d)$ for each of the strands. To begin a description of ideas that go into constructing a stochastic model of the process of nucleotide substitution during meiosis, the strand $s_1^{'}$ will be considered first. As a first approximation and for the sake of simplicity, it will be assumed that the location at which a nucleotide substitution occurs is distributed uniformly on the array of ordered bases $(1, 2, \ldots, d)$. Thus, in any realization of the process with $d \geq 1$ sites, each location in the array $(1, 2, \ldots, d)$ will be chosen with probability $1/d$. In other words, the simulation of the site or location at which a nucleotide substitution occurs is equivalent to drawing a random sample of size 1 from the uniform distribution on the set of positive integers $(1, 2, \ldots, d)$.

As presented in section 6.4, a molecule of DNA is made of bases which are classified as purines and pyrimidines. The purine bases are A - Adenine and G - Guanine, and the pyrimidine bases are C - Cytocine and T - Thymine. In a RNA molecule T is replaced by U - Uracil, which is also a pyrimidine. As stated above, when viewed linearly, a molecule of DNA is made up of pairs of strands such that bases always occur in pairs. These pairs of bases are $A \Leftrightarrow T$ and $G \Longleftrightarrow C$. Observe that these pairs of bonds are such that a purine is always paired with a pyrimidine. Thus, for example, in the pair $A \Leftrightarrow T$ the purine A is bound to the pyrimidine T. Given that a nucleotide substitution occurs in a single strand of DNA, the transition from one base to another will be formulated as a Markov chain with the state space

$$\mathfrak{S} = \ (A, G, C, T) \leftarrow (1, 2, 3, 4)\,. \tag{14.11.1}$$

When writing software to implement this Markov chain, it is convenient to represent the four bases as the integers 1 through 4 as indicated in (14.11,1).

To illustrate the occurrence of a nucleotide substitution consider the DNA molecule $(s_1, s_1^{'})$, where the strand $s_1^{'}$ is complementary to strand s_1 and has arisen as part of the process of doubling the content of DNA in a cell undergoing meiosis. Suppose at some site in the DNA strand s_1 is occupied by the base A, then if there was no nucleotide substitution at this site in strand $s_1^{'}$, the base at this site would be T, but if a nucleotide substitution occurs, then this site in $s_1^{'}$ would be occupied by one of the three bases A, G, C. If the nucleotide substitution $T \rightarrow A$ occurs, then this would be a transition from a pyrimidine T to a purine A. If the substitution $T \rightarrow G$ occurs, then this would again be a transition from a pyrimidine T to a purine G. If the transition $T \rightarrow C$ occurs, however, then this would be a transition from a pyrimidine to a pyrimidine. It is thought that transition from purines to purines and from pyrimidines to pyrimidines occur with higher probabilities than those from purines to pyrimidines and vice versa. As stated in section 6.4, substitutions with the purine class or the pyrimidine are called transitions, but when a substitution occurs among the two classes, it is called a transversion.

Let

$$\boldsymbol{P} = \begin{pmatrix} 0 & p_{12} & p_{13} & p_{14} \\ p_{21} & 0 & p_{23} & p_{24} \\ p_{31} & p_{32} & 0 & p_{34} \\ p_{41} & p_{42} & p_{43} & 0 \end{pmatrix} \tag{14.11.2}$$

denote the transition matrix of the Markov governing the process of nucleotide substitution in the model genome GG under consideration. Then, according to the ordering of the bases in (14.11.1), the transition $A \rightarrow G$ or alternatively $1 \rightarrow 2$, would represent a transition from the purine A to the purine and would thus occur with higher probability than the transitions $1 \rightarrow 3$ and $1 \rightarrow 4$, which represent transition from a purine to a pyrimidine.

Thus, p_{12} would be greater than p_{13} and p_{14}. In the absence of information to the contrary, one could let $p_{13} = p_{14}$ so that from the condition that all the elements of $\boldsymbol{P}$ are ≥ 0 and $p_{12} + p_{13} + p_{14} = 1$, one could deduce that $p_{12} = 1 - 2p_{13}$. This scheme of assigning numerical values to the parameters in the matrix in (14.11.2) would simplify the problem of finding plausible values of the transition probabilities in the transition matrix $\boldsymbol{P}$. The problem of assigning numerical values to the elements of the matrix $\boldsymbol{P}$ have also been expedited the ordering of the bases in (14.11.1). Observe that the first and second bases in this ordering are purines, but the third and fourth bases are pyrimidines. From this observation, it is easy to see that the scheme used in choosing numerical values for the transition probabilities in row 1 of the matrix in (14.11.2) could also be used to assign numerical values to the probabilities in rows 2, 3 and 4. The process just described is also applicable to the pair $(s_2, s_2^{'})$, where it is assumed that nucleotide substitution occur in the stand $s_2^{'}$. For the sake of simplicity, it will also be assumed that in any computer experiment, the numerical values in the Markov transition matrix in (14.11.2) are constant form generation to generation.

The next step in the description of the process of nucleotide substitution that occurs during the doubling of the amount of DNA during meiosis is that of describing a Monte Carlo simulation algorithm to simulate the process described above. As eukaryotes seem to have evolved and editing system that correct copying errors when DNA is replicated, it is thought that a nucleotide substitution occurring during meiosis is a rare event.

Let θ denote the probability that a nucleotide substitution occurs during meiosis. From information on the world wide web and that gained in conversation with workers who studying mutations, it is thought that θ may be somewhere in the interval $(10^{-10}, 10^{-6})$. Let u_1 be a realization of a random variable U that has a uniform distribution on the interval $[0, 1]$. Then, if $u_1 > \theta$, the simulation would proceed that module of the computer simulation in which either crossing over or gene conversion would occur. If $u \leq \theta$, however, then the next set in the simulation procedure would that of simulating a nucleotide substitution. Extensions of these ideas to accommodate other types of mutations will be discussed in a subsequent section.

At this point another simplifying assumption will be made. As suggested above, a nucleotide substitution may occur in either $s_1^{'}$ of the pair $(s_1, s_1^{'})$ or in $s_2^{'}$ of the pair $(s_2, s_2^{'})$. This statement also includes the possibility that a mutation may occur in both the pairs $(s_1, s_1^{'})$ and $(s_2, s_2^{'})$.

This event seems to be unlikely however, so for the sake of simplicity, it will be assumed that the mutation occurs only in the pair $(s_1, s_1^{'})$ or in the pair $(s_2, s_2^{'})$ and each these events occur with probability $1/2 = 0.5$. To simulate this event, let u_2 be another realization of uniform random variable U on the interval $[0, 1]$. Then, if $u_2 \leq 0.5$, the mutation will occur in the pair $(s_1, s_1^{'})$, but if $u_2 > 0.5$, then the mutation occurs in the pair $(s_2, s_2^{'})$.

The next step in the simulation procedure would be that of simulating the site or location where the mutation occurs, and then after the site in chosen, the next step would be that of simulating the base transition. To illustrate these steps in the procedure suppose the nucleotide occurred in the DNA strand $s_1^{'}$ of the pair $(s_1, s_1^{'})$. Then, simulate a random sample of size 1 from a uniform distribution on the set of positive integers $(1, 2, \ldots, d)$. Let k denote the number or site in the set $(1, 2, \ldots, d)$ which was chosen by this procedure. Then, find the base i at site k in the strand s_1. The next step is to select the base $i^{'}$ which is complementary to base i.

In particular, suppose $i^{'}$ is the base 2 as indicated in (14.11.1). Then, choose row 2 of the transition matrix $\boldsymbol{P}$ in (14.11.2). From this row, it can be seen that it is sufficient to consider a three dimensional multinomial distribution with the probability vector $\boldsymbol{p}_2 = (p_{21}, p_{23}, p_{24})$. Then, simulate a random sample of size 1 from this distribution. This sample of size 1 will be one of three indicator vectors; namely $\boldsymbol{\varepsilon}_1 = (1, 0, 0)\,, \boldsymbol{\varepsilon}_2 = (0, 1, 0)$ and $\boldsymbol{\varepsilon}_3 = (0, 0, 1)$. If, for example $\boldsymbol{\varepsilon}_1$ is observed, then the nucleotide substitution $2 \to 1$ would be simulated. The last step in this procedure would be that of replacing 2 with the number 1 at site k in the DNA strand $s_1^{'}$ of the pair $(s_1, s_1^{'})$. After this step is completed, one would proceed to the software module in which the processes of crossing over or gene conversion would be simulated, which have been described in sections 14.9 and 14.10.

To complete this overview of simulating nucleotide substitutions, crossing over and gene conversion during meiosis, it will be helpful to set down the ideas leading to generation of four types of gametes which may be added to the germ lines of a population by some particular genotype under consideration. To begin the symbolic description of this process, let the pair of pairs

$$\left(\left(s_1, s_1^{''}\right), \left(s_2^{''}, s_2\right)\right) \tag{14.11.3}$$

denote a cell during the process of meiosis such that four chromatids are lined up on a spindle prior to the division of the cell into two cells. If the DNA content of this simulated cell is expressed in terms of the number of sites d in each of the four chromatids, then the DNA content of this cell would be $4d$. The strands $s_1^{''}$ and $s_2^{''}$ represent the DNA content of these stands after the processes of nucleotide substitution, crossing over or gene conversion have occurred between non-sister chromatids.

This pair of pairs is random in the sense that it represents a realization of three stochastic processes that occur in the model of meiosis under consideration. During the next stage of meiosis, the cell symbolized in (14.11.3) will divide to yield two cells $(s_1, s_1^{''})$ and $(s_2^{''}, s_2)$ such that the DNA content of each is $2d$. These two cells in turn will divide to produce the set of four gametes $((s_1), (s_1^{''}), (s_2^{''}), (s_2))$.

In this model the gametes (s_1) and (s_2) represent the original parental strands of DNA that have not under gone mutation, crossing over or gene conversion; whereas the gametes $(s_1^{''})$ and $(s_2^{''})$ are those strands of DNA have been changed from the parental strands by the processes of mutation, crossing over or gene conversion. When the three types of processes under consideration are in force, then for any realization of the process of meiosis, only one type of gamete will appear with a probability determined by the values of the parameters taken into in this and the preceding two sections and the genotype of an individual.

To illustrate these ideas in a simple case, suppose that with respect to the markers M_1 and M_2 with their corresponding alleles m_1 and m_2, the genotype of an individual is (M_1M_2, m_1m_2). Next suppose that in the process of meiosis that produced the gametes illustrated above, no nucleotide substitutions occurred with probability $1 - \theta$, and also suppose the process of gene substitution did not occur with probability $1 - \eta$.

Moreover, suppose the process of crossing over among non-sister chromatids did occur with respect to the markers under consideration. Then, to represent the set of possible outcomes of these events symbolically, the set of gametes illustrated above may be written in the form

$$\left((M_1M_2, s_1), \left(M_1m_2, s_1^{''}\right), \left(m_1M_2, s_2^{''}\right), (m_1m_2, s_2)\right).$$

The interpretation of the symbol (M_1M_2, s_1) is, for example, that somewhere in the preassigned order of the bases in the strand of DNA s_1 the bases M_1 and M_2 occur at two sites. Then, in terms of the recombination probability ρ, the gametic distribution corresponding to these four types of gametes is $\boldsymbol{\gamma} = ((1-\rho), \rho/2, \rho/2, (1-\rho)/2)$.

Then, under the assumption that the events just described occur independently, the probability that gamete $(M_1m_2, s_1^{''})$ is produced during meiosis is $(1-\theta) \times (1-\eta) \times \rho/2$. In general, for each genotype under consideration, the gametic distribution of the output of the four types of gametes by the process of meiosis, will be a mixture of distributions that will be discussed in more detail in a subsequent section of this chapter after other mutations that may occur during meiosis have been taken into account.

14.12 Simulating Insertions and Deletions in Evolving DNA Sequences

In a recent paper by Fletcher and Yang (2009), a software package named *INDELible* was described with the purpose of simulating insertions and deletions in evolving sequences of DNA. A focal point of interest for these authors was that, because phylogenetic relationships are rarely known with certainty, computer simulated data are, at present, the best way to characterize the uncertainties surrounding the putative relationships. Briefly, their methods were formulated within a framework of Markov jump processes evolving in continuous evolutionary time, which separates them from the focus of attention in this chapter. Interestingly, the state space of their evolutionary process was the set of all possible configurations of genomic DNA consisting, initially, of some number of base pairs so that evolution of the genome of species under consideration was described in terms of transitions among the states in the state space over time periods consisting of thousand or millions of years evolution. However, in this chapter the development of methods for the Monte Carlo simulation of the evolution of DNA sequences will be on a time scale of generations within a given species and on events occurring during meiosis for every individual in a population. Moreover, the periods of time over which these computer experiments would be conducted are within thousands of generations, rather than millions of years that are often considered in phylogenetic studies. Nevertheless, the distributions used in simulating deletions and insertions as described by these authors, as well as authors cited in the paper, are also useful in simulating the deletions and insertions among individuals on a generational time scale.

Although several distributions for simulating deletions and insertions were discussed by Fletcher and Yang (2009), only two will be discussed here. Let X denote a random variable taking values in the set of non-negative integers $(x \mid x = 0, 1, 2, 3, \ldots)$ and let

$$P\left[X = x\right] = g\left(x\right) \tag{14.12.1}$$

for $x = 0, 1, 2, \ldots$ denoted its probability density function, where $g\left(x\right) \geq 0$ for all x and

$$\sum_{x=0}^{\infty} g\left(x\right) = 1 \tag{14.12.2}$$

Then, by definition, the generating function of this distribution is

$$G\left(s\right) = E\left[s^{X}\right] = \sum_{x=0}^{\infty} g\left(x\right) s^{x} \tag{14.12.3}$$

for $0 \leq s \leq 1$.

Now suppose we expand this generating function in a Maclaurin's series and arrive at the formula

$$G\left(s\right) = G\left(0\right) + G^{(1)}\left(0\right) s + \frac{G^{(2)}\left(0\right)}{2!} s^{2} + \frac{G^{(3)}\left(0\right)}{3!} s^{3} + \cdots . \tag{14.12.4}$$

In this series, the symbol $G^{(x)}\left(0\right)$ is defined by

$$G^{(x)}\left(0\right) = \frac{d^{x} G\left(s\right)}{ds^{x}}, \tag{14.12.5}$$

where the derivative is evaluated at $s = 0$ for $x = 1, 2, \ldots$. By equating coefficients of s^{x} in these two series (14.12.3) and (14.12.4), it follows that

$$g\left(x\right) = \frac{G^{(x)}\left(0\right)}{x!}. \tag{14.12.6}$$

for $x = 1, 2, \ldots$. Thus, if the generating function is given in some explicit functional form, then, in principle, we can deduce a formula for the probability density function of the distribution by using this formula.

Let $p \in (0,1)$, let $q = 1 - p$ and let α be any real number such that $\alpha > 0$. Then, consider the generating function of the form

$$G(s) = \left(\frac{p}{1-qs}\right)^{\alpha}, \tag{14.12.7}$$

which is defined for all $s \in [0,1]$. As we shall see this is a generating function of a probability distribution, but before we deduce the form of density corresponding to this distribution, it will be of interest to make a connection to a well known simple distribution. Consider a random variable X, taking values in the set of non-negative integers, with a probability density function of the form

$$g(x) = pq^{x} \tag{14.12.8}$$

for $x = 0, 1, 2, \ldots$. This function is, of course, one form of the density function for the well known geometric distribution. Its generating function is

$$E\left[s^{X}\right] = \sum_{x=0}^{\infty} pq^{x}s^{x} = \frac{p}{1-qs}. \tag{14.12.9}$$

It is interesting to note that this is a special case of the generating function in (1) with $\alpha = 1$. In fact, it can be shown that when α is a positive integer, then the generating function (14.12.7) is that for the distribution of a sum of α independent geometric random variables with a common density in (14.12.8).

To find the form of the density corresponding to the generating function in (14.12.7), let us expand this function in a Maclaurin's series. Observe that $G(0) = p^{\alpha} = g(0)$ and

$$\frac{dG(s)}{ds} = p^{\alpha}\alpha q (1-qs)^{-(\alpha+1)}. \tag{14.12.10}$$

Therefore,

$$g(1) = \frac{dG(0)}{ds} = p^{\alpha}\alpha q. \tag{14.12.11}$$

In general,

$$\frac{d^{x}G(s)}{ds^{x}} = p^{\alpha}\alpha(\alpha+1)\cdots(\alpha+x-1)\, q^{x}(1-qs)^{-(\alpha+x)}. \tag{14.12.12}$$

Therefore

$$g(x) = p^{\alpha}\frac{\alpha(\alpha+1)\cdots(\alpha+x-1)}{x!}q^{x} \tag{14.12.13}$$

To simplify the notation, let $\alpha^{(x)} = \alpha(\alpha+1)\cdots(\alpha+x-1)$ for $x \geq 1$ and, by definition, let $\alpha^{(0)} = 1$. Then, the density corresponding to the generating function in (14.12.7) may be written in the compact form

$$g(x) = \frac{\alpha^{(x)}}{x!} p^{\alpha} q^{x} \tag{14.12.14}$$

for $x = 0, 1, 2, \ldots$. A random variable X is said to have a negative binomial distribution if it takes values in the set of nonnegative integers and has the density function in (14.12.14). One possible way of thinking about the origin of the term "negative binomial" is to write the generating function in (14.12.7) in the form

$$G(s) = p^{\alpha}(1-qs)^{-\alpha}. \tag{14.12.15}$$

From this form, it is clear that the binomial $(1-qs)$ has the negative exponent $-\alpha$.

There are, of course, many phenomena for which the negative binomial distribution is used to model. For example, when α is a positive integer and p is the probability of the occurrence of some event, then $g(x)$ may be interpreted as the density of the waiting time until α occurrences of the event are observed. In such cases, the coefficient $\alpha^{(x)}/x!$ can be written as an informative combinatorial formula, but, whatever the interpretation of the negative binomial distribution, the density function can always be written in the canonical form (14.12.14).

It can be shown that the expectation of the random variable X is

$$E[X] = \alpha\frac{q}{p}, \tag{14.12.16}$$

and its variance is

$$var[X] = \alpha\frac{q}{p^2}. \tag{14.12.17}$$

It is interesting to note that

$$var[X] = \frac{1}{p}E[X]. \tag{14.12.18}$$

Therefore, because $0 < p < 1$, it follows that $var[X] > E[X]$. Thus, unlike the Poisson distribution for which the expectation and variance are equal, for the case of the negative binomial distribution, the variance is always greater than the expectation.

The formula in (14.12.14) is well suited for the numerical calculation of values of the density function $g(x)$. Let N denote a positive integer such that the set $(0, 1, \ldots, N)$ is the essential support of a negative binomial distribution, given numerical values of the parameters α and p, in the sense that

$$\sum_{x=0}^{N} g(x) \simeq 1. \tag{14.12.19}$$

Then,

$$g_N(x) = \frac{g(x)}{\sum_{x=0}^{N} g(x)} \tag{14.12.20}$$

for $x = 0, 1, \ldots, N$ could be used as a finite approximation to the negative binomial distribution with parameters α and p. In the foregoing chapters of this book an algorithm for simulation realizations of a random variable X_N with the distribution in (14.12.20) was outlined. The output of this algorithm was the set of positive integers $(1, 2, \ldots, N + 1)$, which is suitable for the task at hand, because an insertion or a deletion must consist of at least one base.

As the expectation and variance of random variable X with a negative binomial distribution with parameters α and p are finite for all $\alpha > 0$ and $0 < p < 1$, the distribution does not have heavy tails, which would lead to large realizations of X. Thus, in some cases the negative binomial distribution may not be suitable as model for the distribution of number of bases in deletions or insertions. There is also evidence that the distributions for insertions and deletions may be different. Such considerations have led investigators to consider distributions with heavy tails, see Fletcher and Yang (2009) for more details. One such distribution is known as the power law, which has the probability density function of a random for the variable X of the form

$$g(x) = cx^{-\theta} \tag{14.12.21}$$

for $x = 1, 2, \ldots$, where the parameter θ satisfies the condition $\theta > 1$ and c is a normalizing constant. The constant c is

$$c = \frac{1}{\zeta(\theta)}, \tag{14.12.22}$$

where

$$\zeta(\theta) = \sum_{x=1}^{\infty} x^{-\theta} \tag{14.12.23}$$

is the Riemann Zeta function. The distribution in (14.12.21) is also known as the Zipfian distribution.

Due to the slow convergence of the infinite series defining $\zeta(\theta)$ for some values of θ, the essential support of the distribution may contain a large number of points, which could lead to difficulties when computing realizations of the random variable X. One way of avoiding such difficulties is to consider a continuos analog of the power law, which has a density of the form

$$g(x) = \frac{\theta}{x^{\theta+1}} \tag{14.12.24}$$

for $x \in [1, \infty)$ and $\theta > 0$. In books on mathematical probability and statistics, the density in (14.12.24) is known as the Pareto distribution and is frequently applied as a model for data with outliers, such as very large incomes for some individuals when compared to the population as a whole. The distribution function for the random variable X has the form

$$P[X \leq x] = F(x) = \int_1^x g(s)\, ds = 1 - \frac{1}{x^\theta} \tag{14.12.25}$$

for $x \geq 1$, and the expectation of X is

$$E[X] = \int_1^x s g(s)\, ds = \frac{\theta}{\theta - 1}. \tag{14.12.26}$$

Thus, $E[X]$ exits and is positive and finite if $\theta > 1$. From this formula, it can be seen that the distribution would have heavy tails for those values of θ such that $0 < \theta < 1$ and the expectation $E[X]$ would not exist. It can also be shown that the formula for the variance of this distribution is

$$var[X] = \frac{\theta}{(\theta - 2)(\theta - 1)^2}. \tag{14.12.27}$$

From this formula it can be seen that $var[X]$ is positive and finite if $\theta > 2$. Therefore, for those cases such that $1 < \theta < 2$, the expectation would exist and be finite but the variance would not exist so that in these cases the distribution would also have heavy tails.

An advantage of using this distribution as a candidate for distribution of the number of deletions or insertion is that it is easy to transform and uniform random variable U on the interval $(0, 1)$ to a realization of the random variable X whose range is the set $(1, 2, \ldots)$ of positive integers. Thus, consider the equation

$$U = F(X) = 1 - \frac{1}{X^\theta} \tag{14.12.28}$$

and solve it for X. As a first step note that

$$\frac{1}{X^\theta} = 1 - U. \tag{14.12.29}$$

Thus,

$$X^{\theta} = (1 - U)^{-1}, \tag{14.12.30}$$

which implies

$$X = (1 - U)^{-\theta^{-1}}. \tag{14.12.31}$$

however, because $U = 1 - U$ in distribution, it follows that it suffices to compute

$$X = U^{-\theta^{-1}}. \tag{14.12.32}$$

Finally, by applying the greatest integer function $[\cdot]$, it follows that the range of the function

$$X = \left[U^{-\theta^{-1}}\right] \tag{14.12.33}$$

would be the set of positive integers $(1, 2, \ldots)$. Observe, that if θ is large, θ^{-1} is small so that a realization of X would be "near" 1. On the other hand, if θ is small, then θ^{-1} is large and a realization of X would be "large". If a reader has access to a means for doing arithmetic rapidly, it may be of interest to search for a more precise meaning of these two statements, by choosing numerical values of U and θ.

To fix ideas, consider a computer model of a DNA strand consisting of $c \geq 2$ sites with the set of bases $(b_1, b_2, \ldots, b_c)$. To illustrate the ideas, consider a DNA strand consisting of $c = 5$ sites, which may be represented by the set of integers

$$(1, 2, 3, 4, 5). \tag{14.12.34}$$

Suppose that a segment of DNA consisting of a random number of bases $X \geq 1$ may be inserted at any of the 5 sites in (14.12.34) or attached at either end. By way of an illustration, let 1^* denote a segment of DNA consisting of X bases and suppose it is "attached to base 1. Then, the resulting DNA sequence may be represented in the form $(1^*, b_1, b_2, b_3, b_4, b_5)$ and would consist of $X + 5$ sites or bases. Similar remarks apply to the case a segment of X bases denoted by 5^* were attached to the 5 end of the model DNA strand in (14.12.34). On the other hand, if a DNA strand 2^* consisting of $X \geq 1$ bases were inserted into position 2 in (14.12.34), then the resulting model DNA strand would have the form $(b_1, 2^*, b_3, b_4, b_5)$ and would consist of $X + 4$ bases, which would allow for the deletion of base 2 in the original strand in (14.12.34). However, to illustrate a point, if deletions were not allowed when insertions are under consideration, the simulated

DNA strand would be the sequence $(b_1, b_2, 2^*, b_3, b_4, b_5)$ consisting of $X + 5$ bases and would symbolize the event that the DNA segment 2^* was inserted "between" the bases at sites 2 and 3 of the original DNA stand in (14.12.34).

By ruling out any deletion of bases, for the general case for a model of DNA strand containing c bases there would be $c - 1$ positions for insertions and 2 positions for attachments at either end of the model strand. Therefore, in a simulation model, there would $c - 1 + 2 = c + 1$ positions to consider for simulating insertions and attachments at the end of the model strand. Without the availability any further information, it could be assumed that these $c + 1$ positions are uniformly distributed with a probability density function $f(y) = 1/c + 1$ for $y = 1, 2, \ldots, c + 1$. Let Y denote a random variable with density $f(y)$ and range $y \in (1, 2, \ldots, c.c + 1)$, and let $Y = y$ denote a realization of Y. Similarly, let $X = x$ be a realization of the random variable X, where $x \geq 1$ denotes the number of bases that will inserted into the simulated DNA strand or attached to either end. Then, by definition, if $y = 1$, a segment of DNA consisting of x bases would be attached to the left of end 1 of the DNA strand under consideration. If, for example, $y = 2$, however, then a DNA strand of x bases would be inserted between the bases 1 and 2. The procedure just described would, by convention, also be used in all cases such that $y = 2, \ldots, c$. But, if $y = c + 1$, then a segment of x bases would to attached to the right of position c.

Since in actual simulation insertions and deletions will occur according to some stochastic process as just described, the number c of bases under consideration may vary among individuals in a population. In any experiment, the segment of bases making up an insertion could be sampled at random from a presumed stationary distribution of bases or could be assigned using some plausible assumptions. In an actual experiments, the rationale in choosing the bases in an insertion would need to be made clear.

For the case that only one base is deleted in either or both of the non-sister chromatids during meiosis, the simulation of such event would be straight forward. For example, let c denote the number of bases in a strand of DNA, and let the random variable Y, with range $(1, 2, \ldots, c)$, denote the base or site at which a deletion event occurs. Then, suppose that Y has a uniform distribution on the set $(1, 2, \ldots, c)$ with density function $f(y) = 1/c$ for all $y \in (1, 2, \ldots, c)$. Then, for example, if a realization of $Y = y = 4$, then the base at position 4 would be deleted. On the other hand, let the random variable $X \geq 1$ denote the base or the number of bases that are to be deleted and let $X = x$ denote a realization of the random variable X. Similarly, let $Y = y$ denote a realization of the random variable Y.

As an aid to analyzing the set of events that may occur, suppose $c = 5$ and the sites of simulated DNA strand under consideration are represented as the set of integers

$$(1, 2, 3, 4, 5)\,. \tag{14.12.35}$$

Now suppose the realization of the random variable X is $x = 2$ so that two bases are to be deleted, and suppose the realization of the random Y is $y = 2$. Then, the base at site 2 would be deleted, but at this point it is not clear whether the base on the left or right of the base labeled 2 would also be deleted. If it were assumed that the base to the left of 2 was deleted, then the resulting DNA strand would have the form $(3, 4, 5)$. But, if it were assumed that base to the right of 2 was deleted, then the resulting DNA strand would have the form $(1, 4, 5)$. Next suppose that the realization of the random variable Y is $y = 5$. Then, if it were assumed that the base to the left of site 5 was deleted, the resulting DNA strand would have the form $(1, 2, 3)$. On the other hand, in this case there is no base to the right of base with the label 5. Thus, by convention, only the base 5 would be deleted and the resulting DNA strand would be $(1, 2, 3, 4)$.

One approach to handling the class of situations just outlined, would be to assume that for a realization of $X = x \geq 2$ of the number of bases to be deleted and a realization of the site random variable $Y = y$, the event that $x - 1$ bases to the left of site y were deleted would occur with probability $1/2$. Similarly, the event that $x - 1$ bases to the right of site y were deleted would also be assigned the probability $1/2$. To simplify the situation, an investigator, could, by an agreed upon convention, assume the either $x - 1$ sites to the left of site y or to the right of site y would be deleted with probability one. For the sake of simplicity, assume the deletions are always to the right of site y. If the site y is near the right end of the DNA strand under consideration, other conditions would need to be considered. Let w denote the number of bases to the right of site y. Then, if $w \geq x - 1$, $x - 1$ bases to the right of site y would be deleted. But, if $w < x - 1$, then only w bases would be deleted. In general, let N denote the maximum realization of the random variable X. Then, if N is much smaller than c, the number of bases in the DNA strand under consideration, in a majority of cases the condition $w \geq x - 1$ would be satisfied so that $x - 1$ bases to the right of site y would be deleted, but in the software implementing these ideas, for each realized site $Y = y$ and the number of bases to be deleted $X = x$, the conditions $w \geq x - 1$ or $w < x - 1$ would need to be checked to determine whether $x - 1$ or w bases were to be deleted.

14.13 Simulating Copy Number Variation in an Evolving DNA Sequence

It has been reported that next to single nucleotide polymorphisms, $SNPs$, copy number variants are the most common type of DNA variation that has been observed in the human and other genomes. In humans, copy number variants encompass more DNA than $SNPs$. A copy number variant (CNV) is segment of DNA for which the number of bases vary when the genomes of two or more individuals are compared. The number of bases in these segments can be highly variable and may range from 10 or fewer to as much as a megabase, $1Mb$, or, in some cases, whole genes that may been copied several times. Much information on copy number variation is available on the internet. For example, in one search, when the phrase "copy number variation" was typed into a search engine on the internet, a total of 1,970,000 web sites appeared. Among these sites were published technical papers that were available in the public domain.

The mechanisms underlying $CNVs$ are thought to include genomic rearrangements such as deletions, duplications, inversions and translocations. Methods for simulating deletions were discussed in section 14.12 and, moreover, these methods may also be applied in the simulation of $CNVs$ so that no further consideration to these types of mutations will be given here. Since the preliminary computer experiments being considered in this chapter, suggests that it is unlikely that two or more chromosomes can not be considered in a model genome due to limitations of computer memory. Therefore, the problem of simulating translocations, involving the transfer of segments of DNA from one chromosome to another, will not be considered in this section. However, the simulation of duplications as well as inversions of segments of DNA in a model genome consisting of single strand of DNA will be considered. Before launching into the development of methods to simulate these genomic rearrangements, however, it will be of interest to discuss the significance of $CNVs$ from the points of view of human diseases and evolution.

Copy number variation may be observed by cytogenetic techniques such as fluorescent in situ hybridization, comparative genomic hybridization, array comparative genomic hybridization and virtual karyotyping of SNP

arrays. As a result of these many methods of finding $CNVs$ in genomes, it has been found that such variations are wide spread and are a common phenomenon when the genomes of individuals humans are compared. It has been estimated that 0.4% of the genomes of unrelated people typically differ with respect $CNVs$. It is not surprising, therefore, that these structures have been used as markers in genome wide association studies designed to find regions of DNA that are implicated in diseases of humans and other species. With respect to mental disease in humans, for example, copy number variations has been associated with autism, schizophrenia and idiopathic learning disability. The internet may also be consulted for further information on associations copy number variants with human disease. It is also thought that $CNVs$, such as gene duplication and exon shuffling, may have been predominant mechanisms in driving gene and genome evolution. Recently, it has been observed that copy number variation in intron 1 of $SOX5$ causes pea-comb phenotype in chickens, see Wright *et al.* (2009) for details. This paper is of fundamental interest, because it provides a concrete example of a copy number variant being associated with a phenotype that has selective value in cold climates in that it decreases the risk of frost damage to the comb of a chicken and is thus selected for by breeders.

To some extent the algorithms for simulating the duplication of segments of DNA are similar to those for simulation deletions and insertions developed in section 14.12. Suppose a model DNA strand of c bases is under consideration and let the random variable X with range $(1, 2, 3, \ldots)$, the set of positive integers, denote the random number of bases in a segment of DNA that is to be duplicated, and let this segment with $X = x \geq 1$ bases be denoted by δ. Then, let the random variable Y, whose range is the set of positive integers $(1, 2, \ldots, c)$, denote the site where the duplication begins. For the sake of simplicity, it will be assumed that Y has a uniform distribution on the set of integers $(1, 2, \ldots, c)$. Finally, let the random variable Z, whose range is the set of positive integers, denote the number of times a segment δ of DNA is to be duplicated. Just as in section 14.12, it will be assumed that duplication of the strand δ of DNA proceeds to the right from some starting point in the set $(1, 2, \ldots, c)$.

In some simulation experiment, let $x \geq 1$ denote a realization of the random variable X with a specified distribution, and let the realization $y \in (1, 2, \ldots, c)$ of the random variable Y denote the site where the segment δ with x bases begins. If y is near c, then there may be less than x bases to the right of y so that it would not be possible to find the required number of x bases in the model DNA strand under consideration. The number of bases to the right of site y is $w = c - y$. Thus, x_1, the actual number of bases to be included in the segment δ, would be given by $x_1 = \min(w, x)$. To illustrate the idea of the segment of bases to be duplicated, suppose the realized value of x_1 is 3, and let the (b_y, b_{y+1}, b_{y+2}) denote the set of bases at the sites $y, y+1, y+2$. Then, $\delta = (b_y, b_{y+1}, b_{y+2})$ would be the segment of bases to be duplicated. To find the random number of times the segment δ is to be duplicated, let $z \geq 1$ be a realization of the random variable Z. Then for $z \geq 1$, the number of duplications of the segment δ may be represented in the form $\delta_1, \delta_2, \ldots, \delta_z$. Let B_{y-1} denote the set of bases to the left of site y in the model DNA strand under consideration and let B_{y+3} denote the set of bases to the right of site $y+2$. Then, the modified DNA strand after the event of duplication may be represented in the form $(B_{y-1}, \delta_1, \delta_2, \ldots, \delta_z, B_{y+3})$ and would contain $3z+c$ bases. Like the event of an insertion discussed in section 14.12, a duplication event would increase the number of bases in a model DNA strand, but the occurrence of a duplicated segment of DNA would distinguish a duplication event from an insertion event which would contain no particular structure to the mutated form of a DNA strand.

For the case of a large number of bases that are to be duplicated, such as in the duplication of a gene, it would be necessary for the distribution of the random X to be such that large outliers would occur with some predictable regularity. When a distribution has long tails, such outliers would occur with positive probability. Such a distribution would arise if the value of the parameter θ for the Pareto distribution was in the interval $(1, 2)$. For if $\theta \in (1, 2)$, then the expectation would be finite but the variance would be infinite. So if a random variable X with range $(1, \infty)$ had such a distribution, then the random variable $[X]$ would be a large integer with predictable regularity. On the other hand, if an investigator declined to use the Pareto distribution, then if a random variable W were introduced such that it was uniformly distributed on some finite set S of large integers, then a realization of the random variable W could be used to simulate the number of bases to be duplicated. From these two illustrative examples, a reader can see that there is a multitude of ways to simulate a large number for the number of bases to be duplicated in a simulation experiment.

The event of an inversion of a segment of DNA may be simulated using a modification of the ideas described in the simulation of a duplication event. Let x be a realization of a random variable X denoting the number of bases to be inverted in a model strand of DNA, and let $y \in (1, 2, \ldots, c)$ denote a realization of the random variable Y indicating the site where the inversion began. Again let $x_1 = \min(x, w)$ denote the actual number of bases to be included in the inversion, where w is the number of sites to the right of site y. Then, the segment of DNA to be inverted may be denoted by $\gamma = (b_y, b_{y+1}, \ldots, b_{y+x_1})$. Let $\iota = (b_{y+x_1}, \ldots, b_y)$ denote the inverted segment of DNA, let B_{y-1} denote the set of bases to the left of site y and let B_{y+x_1+1} denote the set of bases to the right of site $y + x_1$. Then, the DNA strand after an inversion event may be represented in the form $(B_{y-1}, \iota, B_{y+x_1+1})$. Observe that in this case, the model DNA strand following the event of a simulated inversion, will still contain c bases. In the next section, a structure designed to simulate the various types of mutational events that may occur during the duplication of the DNA content of a cell during the synthesis stage of meiosis will be developed.

14.14 Simulating Mutational Events and Genetic Recombination During Meiosis

In this section, the probabilities of mutational events that were discussed in the preceding sections of this chapter, which occur during the synthesis stage of meiosis in which the DNA content of a cell is doubled, will be formalized as a set of competing risks. Then, after these competing risks have been formalized in terms of whether mutation events occur or do not occur, attention will be focused on the processes of genetic recombination that may occur by either regular crossing over when the chromosome pairs are aligned on the spindle or through a process called gene conversion or non-reciprocal crossing over. Then, after genetic recombination has been taken into account, attention will be focused on the computation of the gametic distribution and the passing on of gametes to the next generation. The purpose of these exercises is to provide a framework for writing and organizing software designed to simulate events that occur or do not occur during meiosis and whether mutations are or are not passed on to the next generation of offspring.

Let $\theta_{ns}, \theta_d, \theta_{ins}, \theta_{dup}$ and θ_{inv} denote, respectively, the probabilities per meiosis that a nucleotide substitution, a deletion, an insertion, a duplication or an inversion occurs. Then, $\theta_0 = 1 - (\theta_{ns} + \theta_d + \theta_{ins} + \theta_{dup} + \theta_{inv})$ is the probability per meiosis that no mutation occurs during phase of meiosis in which the DNA content of a cell is doubled. Mutations are rare events so the it is plausible that all mutation probabilities belong to the interval of the form $\left(10^{-9}, 10^{-6}\right)$ or even intervals whose lower bound is of the order 10^{-10}. Therefore, θ_0 will be much greater than all the probabilities of the various types of mutation under consideration. Let $\boldsymbol{\theta} = (\theta_0, \theta_{ns}, \theta_d, \theta_{ins}, \theta_{dup}, \theta_{inv})$ denote a six dimensional probability vector and suppose the sample of size 1 is simulated form a multinomial distribution with the probability vector $\boldsymbol{\theta}$. In chapter 5, an algorithm for carrying out this simulation was described. The output of this algorithm was a set $(\epsilon_1, \epsilon_2, \ldots, \epsilon_6)$ of six dimensional indicator vectors, where ϵ_1 is a vector with a 1 in position 1 and 0*s* elsewhere, and indicates the event that no mutation occurred. On the other hand, the vector ϵ_2 has a 1 is position 2 and 0*s* elsewhere and indicates that a nucleotide substitution occurred at some site in the model genome under consideration and the other indicator vectors are defined similarly. If, for example, the vector ϵ_2 was observed, then one would use the algorithm for simulating a nucleotide substitution described in section 14.11. Similar remarks apply to simulating each of the four remaining types of mutation.

Up until now, only one strand of DNA that resulting from a copying process during synthesis phase of meiosis when the DNA content of a cell is doubled has been taken into consideration when formulating models of mutation. However, because during the DNA replication process the two strands of parental DNA split and each of the complementary strands functions as a template for the copying a new strand of DNA, there are actually two sister strands where mutations may occur during the copying process. Because all types of mutations under consideration are thought to be rare events, the probabilities that one or more mutation occur in both new strands of DNA would be very small. Therefore, for the sake of simplicity, mutations will be accounted for only in one strand of DNA which arises as a result of a copying process.

If some type of mating of a female and male were under consideration, then the competing risk process just described would need to carried out for both the female and the male of a mating pair. After taking into account a nucleotide substitution or another type of mutation in either the genome of maternal or paternal parent, one would proceed to implement the algorithms discussed in sections 14.9 and 14.10, where genetic recombination is described. When implementing the ideas developed in this section, it would be necessary to consider η, the probability that a non-reciprocal crossing over event occurs as well as its complement $1 - \eta$, the probability that the normal process of reciprocal crossing over occurs. To simulate these alternative events, one would need to draw an sample of size 1 from a binomial distribution with probability vector $\boldsymbol{\eta} = (1 - \eta, \eta)$ so that the output of this experiment would be two 2-dimensional indicator vectors $\epsilon_1 = (1, 0)$ and $\varepsilon_2 = (0, 1)$. If ϵ_1 were observed, then a reciprocal crossing over event would be simulated, but if the vector ϵ_2 were observed, then a non-reciprocal crossover event would be simulated. In either case, one would then proceed to compute the gametic distribution for each parent and draw a sample of size one from each parent's gametic distribution to simulate the genotype of an offspring. It should be noted that the gamete contributed by either parent may or may not contain a simulated mutation. If either of the parent's gametes contributed to an offspring contain a mutation, then this mutation will be passed on to the next generation.

The set of algorithms just described is applicable to a diploid species, such as man, in which two sexes, females and males, form various combinations of partnerships to produce the next generation of offspring. To implement and execute these algorithms for populations consisting of hundreds or thousands of individuals would result in a computer intensive experimentation in which each experiment may take a long span of time to complete. Rather than launching into this more complicated case, it would seem advisable to consider the genome of a species, whose genome was small in comparison with the human genome and that of other mammalian species. There are species of bacteria or structures in cells such as mitochondria, which are thought to have evolved from a species of bacteria, with relatively small genomes, which could serve as model genomes for computer simulation experiments.

For example, the genome of Human mitochondrion consists of 16,569 base pairs and contains 37 protein coding genes. Mitochondria reproduce asexually and do not under go the process of meiosis so that genetic recombination would not need to be considered and attention could be focused mutations that arise as a resulting of errors in the DNA copying process during cell division. As discussed in chapters 6, 7 and 8, the simulation of mutations in the genome of the Human mitochondria would of interest in its own right. In this connection, it should be recalled, that in these chapters the only type of mutation that was taken into account was nucleotide substitution, but by using the algorithms presented in this chapter other mutations such as deletions, insertions, duplications and inversions could also be taken into account.

Another organism whose genome could serve as a model in computer simulation experiments would be that of Mycoplasma genitalium, whose genome consists of 580,073 base pairs and contains 485 protein coding genes. This genome was sequenced by scientists at the J. Craig Venter Institute and this work was followed up by systematically destroying genes by insertion mutations to determine which ones were essential to life. Of the 485 protein coding genes, it was found that only 381 were essential to life. More recently, the genome of this species has been constructed from the DNA of a yeast genome, which was, in a sense, a successful attempt to create life. More details on these experiments may be found on the internet. In principle, computer experiments on the short term evolution of a model genome of this species in terms of the types mutation discussed in this section could be carried out and compared with changes in the genome in generations of cultures of this organism. Such experiments would be of great interest, because they would provide a basis for checking the validity of the assumptions underlying the mutational process discussed in this chapter. For example, would the assumption of a uniform distribution for the sites at which a mutation occurred be consistent with the observed experimental data?

Bibliography

[1] Akey, J., Zhang, G., Zhang, K., Jun, L. and Shriver, M. D. (2002) Interrogating a high-density map for signatures of natural selection. Genome Research **12**:1805–1814, Cold Spring Harbor Laboratory Press.

[2] Altshuler, D. and Clark, A. G. (2005) Harvesting medical Information from the human family tree. Science **307**:1052–1053.

[3] Balloux F. and Lugon-Moulini, N. (2002) The Estimation of population differentiation with microsatellite markers. Molecular Ecology **11**:155–165.

[4] Benjamini, Y. and Hochberg, Y. (1995) Controlling false discovery rate: A practical and powerful approach to multiple testing. J. Roy. Soc. Ser. B **57**:289–300.

[5] Benjamini, Y. and Yekutieli, D. (2001) The control of false discovery rate under dependency. Ann. Statist. **29**:1165–1188.

[6] Benjamini, Y. and Yekutieli, D. (2005) False discovery rate adjusted multiple confidence intervals for selected parameters. J. Amer. Statist. Assoc. **100**:71–93.

[7] Chadeau-Hyam, M. *et al.* (2008) Fregene: Simulation of realistic sequence-level-data in populations and ascertained samples. BMC Bioinformatics **9**:364.

[8] Cochran, G. and Harpending, H. (2009) **The 10,000 Year Explosion - How Civilization Accelerated Evolution**. Basic Books. New York.

[9] Deely, J. J. and Lindley, D. V. (1981) Bayes empirical Bayes. J. Am. Stat. Assoc. **76**:833–841.

[10] Devlin, B. and Risch, N. (1995) A comparson of linkage disequilibrium measures for fine-scale mapping. Genomics **29**:311–322.

[11] Durrett, R. (2008) **Probability Models for DNA Sequence Evolution, Second Edition.** Springer Science.

[12] Efron, B. (2008) Microarrays, Empirical Bayes and the Two-Groups Model. Statistical Science **23**:1–22.

[13] Excoffier, L. (2003) Analysis of population subdivision. (In Handbook of Statistical Genetics, 2nd Edition. Edited by D. J. Balding, M. Bishop and C. Cannings) pp. 713–750.

[14] Fletcher, W. and Yang, Z. (2009) INDELible: A flexible simulator of biological sequence evolution. Mol. Biol. Evol. **26**:1879–1888.

[15] Fry, A. E. *et al.* (2006) Haplotype homozygosity and derived alleles in the human genome. Am. J. Hum. Genet. **78**:1053–1059.

[16] Gardner, M. J. *et al.* (2002) Genome sequence of the human malaria parasite Plasmodium falcipaum. Nature **419**:498–511.

[17] Goldman, N. and Yang, Z. (1994) A codon based model for nucleotide substitution for protein-coding DNA sequences. Molecular Biology and Evolution **11**:725–736.

[18] Grossman, S. R. *et al.* (2010) A composite of multiple signals distinguishes causal variants in regions of positive selection. Science **327**:883–886.

[19] Hinds, D. *et al.* (2005) Whole-genome patterns of common DNA variation in three human populations. Science **307**:1072–1053.

[20] Hoggart, C. J. *et al.* (2007) Sequence level population simulations over large genomic regions. Genetics **177**:1725–1731.

[21] Hoggart, C. J. *et al.* (2008) Genome-wide significance for dense SNP and resequencing data. Genetic Epidemiology **32**:179–185.

[22] Hudson, R. R. (2002) Generating samples under a Wright-Fisher neutral model. Bioinformatics **18**:337–338.

[23] Kass, R. E. and Raferty, A. E. (1993) Bayes factors and model uncertainty. Technical Report No. 254, Department of Statistics, Univ. of Washington, Seattle, Washington 98195.

[24] Li, W. H. (1997) **Molecular Evolution**. Sinauer Associates Inc. Sunderland, Mass 01375.

[25] Lewontin, R. and Krakauer, J. (1973) Distribution of gene frequencies as a test of the theory of selective neutrality of polymorphisms. Genetics **74**:175–195.

[26] Mailund, T. M. H. *et al.* (2005) $Coa\,\mathrm{Si}\,m$: a flexible environment for simulating genetic data under coalescent models. BMC Bioinformatics **6**:252.

[27] Marjoram, P. and Wall, J. D. (2006) Fast coalescent simulation. BMC Bioinformatics **7**:16.

[28] Nicholson, G. *et al.* (2002) Assessing population differentiation and isolation from single-nucleotide polymorphism data. J. R. Statist. Soc. B **64**:695–715.

[29] Nielsen, R. *et al.* (2005) A scan for positively selected genes in genomes of humans and chimpanzees. PLoS Biology 3:e170,10.1371/journal.pbio.0030170.

[30] Peng, B. and Kimmel, M. (2005) *simuPOP* : a forward in time population genetics simulation environment. Bioinformatics **21**:3686-3687.

[31] Peng, B. *et al.* (2007) Forward in time simulations of human populations with complex diseases. *PLoS* Genetics **3**:407–420.

[32] Pritchard, J. and Przeworski, M. (2001) Linkage disequilibrium in humans: models and data. America Journal of Human Genetics **69**:1–14.

[33] Raj, T. (2009) Molecular Signatures of Natural and Artificial Selection in Mammalian Genomes. Ph.D. Thesis, St. John's College, University of Cambridge, UK.

[34] Schaffner, F. *et al.* (2005) Calibrating a coalescent simulation of human genome sequence simulation. Genome Research **15**:1576–1583.

[35] Sabeti, P. C. *et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature **419**: 832–837.

[36] Sabeti, P. C. *et al.* (2006) Positive selection in the human linage. Science **312**:1614–1620.

[37] Spencer, C. C. A. and Coop, G. (2004) *Sel*Si*m* : a program to simulate population genetic data with natural selection and recombination. Bioinformatics **20**:3373–3375.

[38] Su, X. *et al.* (1999) A genetic map of recombination parameters of the human malaria parasite, Plasmodium falciparum. Science **286**:1351–1353.

[39] Voight, B. F. *et al.* (2006) A map of recent positive selection in the human genome. PLoS Biology **4**:446–458.

[40] Weir, B. S. and Hill, W. G. (2002) Estimating F-Statistics. Annu. Rev. Genet. **36**:721–750.

[41] Wells, S. (2002) **The Journey of Man - A Genetic Odyssey**. Princeton University Press, Princeton, Oxford.

[42] Wells, S. (2006) **Deep Ancestry - Inside the Genographic Project**. National Geographic Society, Washington, D. C., U.S.A.

[43] Wells, S. (2009) ***DVD* The Human Family Tree - Tracing the Human Journey Through Time**. National Geographic Society, Washington, D. C., U.S.A.

[44] Wright, D. *et al.* (2009) Copy number variation in Intron 1 of *SOX*5 causes the Pea-comb phenotype in chickens. PLoS Genetics 5, Issue 6, e1000512.

[45] Yang, Z. (1997) *PAML* : A program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13**:555–556.

[46] Yang, Z. (1998) Likelihood ratio tests for positive selection and application to primate lysozyme evolution. Molecular Biology and Evolution. **15**:568–573.

[47] Yang, Z. (2006) **Computational Molecular Biology**. Oxford University Press, Oxford, New York.

[48] Yang, Z. (2007) *PAML* 4: Phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution. **24**:1586–1591.

[49] Yang, Z. and Bielawski, J. P. (2000) Statistical methods for detecting molecular adaptation. Trends in Ecology and Evolution. **15**:496–503.

[50] Yang, Z. and Nielsen, R. (2002) Codon substitution models for detecting molecular adaptation at individual sites along specific lineages. Molecular Biology and Evolution. **19**:908–917.

[51] Yang, Z. *et al.* (2000) Codon substitution models for heterogeneous selection pressure at amino acid sites. Genetics. **155**:431–449.

[52] Yang, Z. *et al.* (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. Molecular Biology and Evolution. **22**:1107–1118.

[53] Zhao, H. and Speed, T. P. (1996) On genetic map functions. Genetics **142**:1369–1377.

Chapter 15

Suggestions for Further Research, Reading and Viewing

15.1 Introduction

In this short chapter, devoted to suggestions for continuing research, further reading and the viewing *DVDs* on genetics and evolution, only three substantive sections were included. In section 15.2, suggestions for the further development of the material presented in chapters 9,10,11 and 12 are outlined. Perhaps, among these suggestions, the two that will bear the most fruit concern the proposal for extending methods for simulating genealogies from the one type to the multitype case. When this extension is accomplished, it will be possible to estimate the distribution of the age of a mutation and whether it appeared in more than one individual. A second suggestion concerns further development of the material presented in chapter 12 on the evolution in age structured populations. When the ideas put forth in this chapter are more fully developed to include Monte Carlo simulation methods, they will provide a medium for the study of the interactions genetics and the development of culture in age structured populations. Such populations provide a milieu for the transmission of culture among cohorts of age groups and an assessment of its uncertainties in this process may be made by using Monte Carlo simulation methods. These simulation experiments will not only be useful in providing confirmatory evidence for ideas concerning the joint evolution of humans and culture but they may also suggest additional ideas concerning this joint evolution, and, at the same time, they will take into account that human evolution proceeds on a much slower time scale than cultural evolution.

Section 15.3 is devoted to suggestions for further theoretical development of methods concerning the evolution of *DNA* and the development of Monte Carlo simulation methods to study the process introduced in chapter

14 by computer simulation. Due to the complexity of the issues that arise when considering the evolution of DNA through the processes of various types of mutation and selection, further efforts in this regard necessitates the organization of teams of interacting individuals. Some members of the team, for example, will be concerned with the development of the mathematics and software underlying the methods and maintaining transparency; while others will be concerned with the problems of relating the developing mathematics and software to description and analysis of existing data in either haplotype form or sequenced genomes of individuals.

An interesting question that arises when considering the development of stochastic methods to simulate the evolution of DNA is that of embedding a deterministic model in a stochastic process. At the moment, it is not clear what form this procedure may take when evolution at the molecular level is under consideration which involves sites of a DNA molecule as compared to estimating counts of numbers of individuals of each type in an evolving population at the macro level. It is suggested in section 15.3 that research efforts on developing methods for the simulating the evolution of DNA begin with genomes of microorganisms with relatively small genomes before proceeding to eukaryotes with more complex genomes and the presence of genetic recombination. It appears that at this juncture in time, that the full integration of methods of simulating the evolution of DNA with the age structured models of population introduced in chapter 12 will occur only at sometime in the distant future.

The list of references for further reading and study presented in section 15.4 is short, and, it seems very likely that readers will add to this list to suit their particular interests.

15.2 Suggestions for Further Research on Self-Regulating Branching Processes

Essentially all the material on the applications of self regulating branching processes to genetics and evolution as presented in chapter 9,10,11 and 12 is in a preliminary state and would profit from new research initiatives and directions within the branching processes and Mendelian paradigms. With little effort, it would be possible to set down a long list of initiatives, but in what follows attention will be limited to only a few suggestions that seem to hold the most promise for future research.

One of the most interesting and useful extensions to the techniques of simulating genealogies for the one type of individual case developed in chapter 9 would be the extension to the cases in which individuals of two or more types were accommodated in the model. Such structures would allow for the possibility of estimating the distribution of the age of a mutation from individuals sampled in some generation in a simulated multitype genealogy and whether a mutation arose more than once in two or more individuals during the evolution of a simulated population.

With reference to the two sex multitype branching process introduced in chapter 11 in which the evolution of a populations was on a time scale of discrete generations, perhaps the most useful and interesting extension of this formulation would be that of parameterizing the acceptance probabilities so that two or more Mendelian loci with two or more alleles could be included in the formulation. Given such an extension, it would be possible to study approaches to linkage equilibrium in populations evolving stochastically from a small founder populations under various assumption such as whether the mating was random, there was or was not mutation and whether selection was present or absent. Given these results on the convergence to linkage equilibrium in finite population, it would be possible to further assess whether evidence for linkage disequilibrium was a reliable signal for the action of natural selection.

From the point of view of developing a framework for simulation human evolution and the coevolution of agriculture and the domestication of plants, fungi and animals during the last 10,000 or more years, the age-dependent two sex and self regulating branching process introduced in chapter 12 is the most relevant. In a word, this class of branching processes belongs to an algorithmic extension of what is known as the general multitype branching process. When this stochastic structure is further developed, it will provide a framework for the evolution of culture in human populations by including overlapping generations that provide a social structure for the continuing development of culture in which evolving parental generations pass on this culture to their offspring who in turn continue the process of development and education from generation to generation.

It is thought by many that this development of culture has genetic components on which natural selection has acted as humans evolved in an environment which was, to some extent, of their own creation, through the domestication of plants, fungi, (yeasts) and animals that supply food, shelter, clothing, entertainment and companionship. After providing a preliminary formulation of such an age dependent process in chapter 12, software was written to implement the embedded deterministic model for a demonstration that software could, indeed, be written for the computer implementation of the model.

To further develop a formulation to accommodate the description outlined above, would require a need for a parametrization of the module describing partnership formation and the resulting production of offspring as well as the writing of software to compute samples of Monte Carlo realizations of the process and provide statistically informative summarizations of the simulated data. In this connection, the most difficult problem which needs to be solved in designing and writing of the software is to accommodate the age structure of the simulated data as well as methods to produce statistically informative summarizations of data on an evolving age structured population.

15.3 Suggestions for Continuing Development of Stochastic Models of Genomic Evolution

Just as was the case for further developed of self regulating branching process discussed in chapters 9,10,11 and 12, many possible line of future research could also be discussed to further the development model of genomic evolution discussed in chapter 14. However, in this section attention will be focused on only a few tentative research initiatives.

As suggested in chapter 14, rather than starting with complex diploid genomes, which are made up of billions of base pairs, it seem advisable to start with simpler organism with much smaller genomes such as that of Mycoplasma genitalium, which has been reported to be made of up 580,073 base pairs and 485 genes. There is an extensive literature on the genome of this species of bacteria as can be seen if the phrase "M. genitalium genome" is typed into a search engine for the internet. A primary reason for the wide interest in this species of bacteria is that it is the cause of a sexually transmitted disease in humans and has thus been given much attention in the medical literature. One among many papers and other references about its genome is that of Suthers, P. F. *et al.* (2009) in which a genome-scale metabolic reconstruction of the M. genitalium genome is reported. Interestingly, these authors report that there are approximately 480 protein coding genes in this organism. Some of the methods used by these authors may be of interest when construction model genomes for computer simulation experiments.

As this organism seems to be cultured routinely in connection with the diagnosis of sexually transmitted diseases, it seems likely that it could also be used as a model to study the evolution of populations in a laboratory. By way of illustrating a potential research initiative, suppose it was also possible to use a computer model of this genome to design Monte Carlo simulation experiments to study mutation and selection in this organism that could, in principle, be checked experimentally. A class of self regulating branching processes that would be suitable as a model to study the evolution of an organism with a genome of about 480 kilo bases would be the class of multitype branching processes discussed in chapter 10. Initially, it may be advisable to limit mutations only to nucleotide substitutions, but as time passes deletions, insertion, duplications and inversions could also be included in the study.

In principle, one could study how these types of mutation would affect the 480 protein coding genes of this organism and how these mutations affect the two components of natural selection included in the stochastic processes discussed in chapter 10. Recall that these components were the expected number of offspring produced per individual per generation and the ability of each individual of a given type to survive and produce offspring. As bacteria reproduce by binary fission, the number of offspring produced by each individual is two so that this component of natural selection would not be operative in such populations. However, the ability to survive to produce offspring would be in force and would, therefore, be the driving force in the evolution of such populations. It should also be recalled, that the ability to survive and reproduce per generation may depend on population size in the class of processes discussed in chapter 10 and this type of natural selection may also be present in cultured experimental populations of M. genitalium.

The research initiative just discussed also suggests that there are many applications for models of genome evolution in those organism that cause diseases in man. In this connection, two strains, or species, of E. coli. are of interest. One strain has the title E. coli. K-12 which has a genome of about 4,639,221 base pairs and 4,377 genes. Evidently, 4,290 of these genes code for proteins, while the rest code for $RNAs$. At this point, it would be of interest to recall that RNA coding genes fit the definition of a gene presented in chapter 13. A second strain of E. coli. O157:H7 is pathogenic in humans and has a genome of about 5.44×10^6 base pairs and 5,416 genes. Among these genes, 1,346 are not found in E. coli. K-12. An interesting question is: if E. coli. K-12 is viewed as the ancestral species, then what were the evolutionary steps at the genomic level that lead to the emergence of E. coli. O157:H7? In a comparative genomic study of these two species of E. coli., it would be of great interest to study the differences in the genomes of these two species. For example, what sort of mutational events occurred in the evolution of these to species? In other words, were deletions, insertions, duplications and inversions among the mutational events separating these species? One could, of course, also entertain the notion that E. coli. O157:H7 was the ancestral species and the processes of mutation and natural selection gave rise to the non-pathogenic form E. coli. K-12.

If the phrase "Genome Size" is entered into a search engine on the internet, additional information on the sizes of genomes may be obtained along with the number of genes and some information on their functions. An example of such a search is the fungus Neurospora crassa, which is also known as bread mold. This eukaryotic organism is a central organism in the history of twentieth century genetics, biochemistry and molecular biology. In 2003, Galagen *et al.* (2003) reported on a high-quality draft sequence of the N. crassa genome. These authors state that this genome is made up approximately 40 megabases and encodes for about 10,000 protein-coding genes, which is only about 25% fewer than the fruit fly Drosophila melanogaster. Neurospora possesses the widest array of genome defense mechanisms known for any eukaryotic organism, which includes a process unique to fungi called repeat-induced point mutation (RIP). An analysis of this genome suggests that RIP has had a profound impact on genome evolution and has slowed the creation of new genes through genomic duplication, which resulted in a genome with an unusually low proportion of closely related genes.

If an investigator were confined to a desk top computer for conducting experiments on the evolution of a genome involving mutation and selection, it is unlikely that 40 megabase genomes could be accommodated, given the memory capacity of present day desk top computers. However, if a network of computers were available, it may be possible to accommodate 40 megabase genomes with existing technology. Rather than attempting to launch into experiments with model genomes as large as 40 megabases, it seems advisable to start with a model genome consisting of about 500,000 base pairs and let this serve a baseline for writing software to implement the preliminary models of genetic recombination and mutation discussed in chapter 14.

It also seem prudent to start with a species that reproduces by binary fission so the complications of linkage and genetic recombination could initially be ignored. In such a system, attention could be focused on developing software modules for the mutational processes of nucleotide substitution, deletions, insertions, duplications and inversions. If some workable cooperative arrangement could be worked out with other investigators such that simulated mutations in a model genome could be compared with sequenced genomes of strains of bacteria such as M. genitalium, one could assess whether the assumptions used in the models underlying mutational process were reasonable when compared with data. One such a assumption is that of assuming the sites where mutations occur or begin are uniformly distributed on a set of sites or does the sequenced data suggest that mutations seem to be concentrated some regions of a genome more than others. In principle, other assumptions underlying the models of mutational processes could also be checked for validity.

After the performance software modules that implement the types of mutations under consideration were tested, an investigator could begin the process of extending the software to accommodate diploid populations and the processes of genetic recombination, which would include crossing over during meiosis as well as gene conversion in a model genome of about 500,000 base pairs. One of the challenges in developing software in the diploid case would be that of developing module for keeping track of the types of mutations and genetic recombinations that occur among the individuals in a simulated population in each generation. Of course, such problems also arise when considering a haploid species that reproduced by binary fission. Whenever actual sequenced data are available, the simulated data could be compared with the observed data as a check on the validity of the assumptions underlying the models for mutation and genetic recombination. As the availability of sequenced human genomes increases as the personal genome project develops, investigators will have further opportunities for comparing simulated data on a model genome with the variability observed in actual sequenced data.

15.4 A Brief List of References on Genetics and Evolution for Further Study

The primary source of inspiration that led to the writing of this book did not come from studying the literature on stochastic processes, probability and statistics but rather from reading books and viewing *DVDs* on biology and evolution. Given this inspirational background, the techniques used to write this book, which were based on a working knowledge of stochastic processes and statistics along with a minimal competence in computer science, merely became a set of tools to be used in formulating and implementing computer simulation models to carry out exploratory Monte Carlo simulation experiments on the quantification of mutation and selection, which have been widely recognized as among the important driving forces of evolution. The books that served as sources inspiration were not written just for an audience of scientists but rather for general readers with interests in science and evolution. Interestingly, such books are often also of great interest to scientists and are sources of inspiration and pleasant diversions from the dryer technical books and papers intended only for an audience of their peers.

For the case of the senior author, who in 1956 earned a Ph.D. in genetics but had not kept up with the vast literature on this subject, due diversion into intensive studies of probability and statistics at a professional level in departments of mathematics and statistics, the reading of the books and the viewing of $DVDs$ listed below was a pleasant home-coming to genetics and evolution after spending at least four decades in other disciplines. Among the first book on the reading list was that of Carroll (2005) on endless forms must beautiful and the new science, Evo Devo, which combines the studies of development and evolution. As a result of reading of this book, a keener awareness of regulation of genes was attained. A second book by the same author, Carroll (2006), on the making of the fittest and DNA and the ultimate forensic record of evolution, further strengthened the ideas that evidence for evolution could be discerned by examining the DNA of among individuals of a species and among species. A third influential book on the reading list was that of Coyne (2009) with the title “Why Evolution is True”. In science one uses the word, true, only rarely due the tentative nature and current scientific theories, which may be over thrown when new evidence becomes available. But for Coyne the use of the word, true, is justified because evolution is more than just a theory, for many of its predictions can be verified by the fossil record or by other evidence. It should emphasized however, that such assertions are to some extent still tentative due to statistical and other uncertainties. Coyne along with other authors have made convincing arguments that the facts of evolution based on evidence that has been debated in the public square provides a firmer bases for the teaching evolution in high schools than those theories that are based solely on religious convictions such as intelligent design for which many convincing counter examples in biology may be found, see Coyne (2009) for details.

Up until recently, it was thought by many, including the senior author, that man had not evolved significantly during the last 10,000 years as civilization based on agriculture developed. However, when reading the fourth book on the reading list by Cochran and Harpending (2009) with the title "The 10,000 Year Explosion - How Civilization Accelerated Human Evolution" doubts arose as to whether the idea that man had not evolved significantly during the last 10,000 years was tenable. After performing the computer experiments with a version of a two sex model as reported in chapter 11, in which was shown that a recessive gene could rise to predominance in a population due the to process of sexual selection within 10,000 or fewer years, there were also doubts regarding the idea that man had not changed genetically during geologically recent times even before the book by Cochran and Harpending (2009) had been read. In subsequent searches of the literature and the internet, further evidence was found that was contrary to the idea that man evolved little during the last 10,000 years. Among this evidence were populations who depended on milk from cows in which a gene for lactose tolerance had become prevalent. Other evidence suggested that some types of cognitive abilities had also arisen in man due to complexity on human networks that inevitably arise as total population size increases in civilized communities.

Some of "The Great Courses" from the teaching company presented in a DVD format have also been very helpful in forming the ideas underlying this book. One such course was on the origins of life by Hazen (2005) in which such topics as definition of life, self replicating molecular systems, the RNA world, the pre-RNA world, natural selection and competition among other topics are presented. A second course on genetics presented by Sadava (2008) was a nice refresher of ideas in genetics for someone who had not heard genetics verbalized for nearly 40 years. The title of this course was "Understanding Genetics: DNA, Genes, and Their Real-World Applications" and the lectures on DNA were very informative from both the historical and substantive points of view. Another course on genetics was that of Silver (2009) with the title "The Science of Self". This former physicist was particularly adept at presenting material in informative graphic forms and provided an interesting account of the use of biotechnology and microarrays that are now commonly used in a variety types of genetic research. Another topic that is of great interest, particularly in human evolution, is the biology of human behavior. A set of lectures by Sapolsky (2005) with the title "Biology and Human Behavior: The Neurological Origins of Individuality" were vary informative form the standpoint of how genes influence the brain through the processes of transcription and translation in the neurons, and their consequent effects on our behavior. Of particular interest is that this author raises the concept of the interactions of genotypes and their environments to new heights in the analysis of human behavior and that of other animals. Finally, the course by Sojka (2009) with the title "Understanding the Human Factor: Life and Its Impact" provides an interesting account of the development of agriculture and domestication of animals, plants, fungi and other life forms that humans now rely on for sustenance. In these lectures clear evidence is presented that man has greatly influence the evolution of his domesticates, which also strongly suggests that his domesticates have also influenced the evolution of man during the last 10,000 years by accommodating genetically to consuming milk products and also the negative effects of some proteins in wheat flower to which some people are allergic.

DVDs of *NOVA* programs on science, which are shown by a network of PBS (Public Broadcasting Service) television channels were also very useful sources of information. Two of these programs shown in connection with the 200-th anniversary of Charles Darwin's birth have been viewed by millions and also by groups interested in continuing education. One *DVD* consisting three one-hour episodes has the title "Becoming Human - Unearthing our Earliest Ancestors" and traces the evolution of a putative human ancestor of going back 4,000,000 years of more to the emergence of Homo erectus, who migrated out of Africa into Europe and Asia, and is thought to have had many human-like traits. A final episode is devoted to the emergence on modern man, and, among other factors, it is thought that a severe drought in Africa due to a very cold glaciation period drove some of ancestors the southern coast of Africa, where abundant species of shell fish were available for food along with plants in whose roots were stored rich carbohydrates that were a source of energy. There is also evidence that these early modern humans had advanced cognitive abilities in that they new the timing of tides that made it possible in some seasons of the year to collect shell fish in tidal areas that were rich in protein. A second *DVD* consisting on two one-hour episodes has the title "What Darwin Never Knew" and provides a rich sample of evidence for evolution, based on the applications of technology that was not available during his day but lends convincing support of his theories of evolution by natural selection. Also contained in this *DVD* are examples of discoveries from modern medical research on inherited disease that are best explained within an evolutionary paradigm.

There are also two other books that were read in preparation for the writing of this book. One of these books was an autobiography of J.C. Venter with the title “A Life Decoded - My Genome: My Life”, Venter (2007). Venter headed up the private effort to sequence the human genome by what was called the shot gun method, which involved an effort to solve a jig saw puzzle consisting of billions of pieces, with the help of sophisticated discrete mathematics and software implemented on computers specifically designed for the task. It is also a personal account, among other things, of positive and negative interactions with peers, which all scientists experience in one form or another throughout their careers. A second book dealt with the human tendency for religiosity and what is thought by many that this tendency provides a bases for conflicts between religion and scientific theories of evolution. Michael Dowd, who is known as America’s evolutionary evangelist, meets the supposed conflict head on in an interesting book with the title “Thank God for Evolution - How the Marriage of Science and Religion will Transform Your Life and Our World”, Dowd (2007), which has been well received by members of the scientific and faith communities as well as various assortments of atheists, agnostics and free thinkers. Perhaps the most compelling reason for the good reception of this book is that many people in various camps with differing value systems find that declaring an end to the often unseemly and unnecessary war between religion and science would be a welcomed event that is long over due.

Bibliography

[1] Carroll, S. B. (2005) **Endless Forms Most Beautiful - The New Science of Evo Devo**. W. W. Norton & Company, London and New York.

[2] Carroll, S. B. (2006) **The Making of the Fittest - *DNA* and the Ultimate Forensic Record of Evolution**. W. W. Norton & Company, London and New York.

[3] Cochran, G. and Harpending, H. (2009) **The 10,000 Year Explosion - How Civilization Accelerated Evolution**. Basic Books. New York.

[4] Coyne, J. A. (2009) **Why Evolution is True**. Viking - Penguin Group, New York, London, Toronto, Dublin.

[5] Dowd, M. (2007) **Thank God for Evolution - How the Marriage of Science and Religion will Transform Your Life and Our World**. Viking - Penguin Group, New York, London, Toronto, Dublin. Viking - Penguin Group, New York, London, Toronto, Dublin.

[6] Galagan, J. E. *et al.* (2003) The genome sequence of the filamentous fungus Neurospora crassa Nature Apr. 24;**422**(6934):859—868.

[7] Hazen, R. M. (2005) **Origins of Life**. The Teaching Company. Chantilly, Virginia, USA. www.teach12.com.

[8] **DVD: Becoming Human - Unearthing our Earliest Ancestors** shopPBS.org.

[9] **DNA: What Darwin Never Knew**. shopPBS.org.

[10] Sadava, D. (2008) **Understanding Genetics:. *DNA*, Genes and Their Real-World Applications**. The Teaching Company. Chantilly, Virginia, USA. www.teach12.com.

[11] Sapolsky, R. (2005) **Biology and Human Behavior: The Neurological Origins of Individuality**. The Teaching Company. Chantilly, Virginia, USA. www.teach12.com.

[12] Silver, L. M. (2009) **The Science of Self**. The Teaching Company. Chantilly, Virginia, USA. www.teach12.com.

[13] Sojka, G. A. (2009) **Understanding the Human Factor: Life and its Impact**. The Teaching Company. Chantilly, Virginia, USA. www.teach12.com.

[14] Suthers, P. K. *et al.* (2009) A Genome-Scale Metabolic Reconstruction of Mycoplasma genitalium, iPS189. PLoS Computational Biology vol. 5, issue 2, e1000285.

[15] Venter, J. C. (2007) **A Life Decoded - My Genome: My Life** Penguin Group, New York, London, Toronto, Dublin.

Index